

NAG2-899

3E 2204

QUADRATURE, INTERPOLATION AND
OBSERVABILITY

by

LUCILLE McDANIEL HODGES, B.A., M.A.

A DISSERTATION


IN


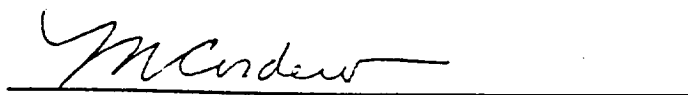
MATHEMATICS

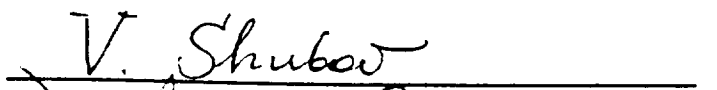

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY

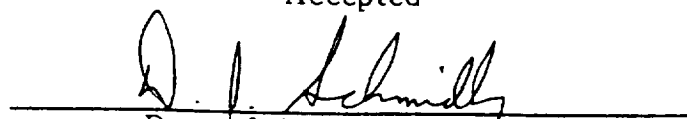
Approved


Chairperson of the Committee

Accepted


Dean of the Graduate School

December, 1997

CONTENTS

ABSTRACT	iii
I. INTRODUCTION	1
II. ORTHOGONAL FUNCTIONS	5
III. GAUSSIAN QUADRATURE	10
IV. STABILITY	14
V. OBSERVABILITY	17
VI. TCHEBYCHEFF SYSTEMS	22
VII. EXPONENTIAL INTERPOLATION	52
VIII. QUADRATURE METHODS	58
8.1 Nilpotent Matrices	60
8.2 Matrices With One Real Eigenvalue	60
8.3 Distinct Real Eigenvalues	63
8.4 Real Repeated Eigenvalues	65
8.5 Complex Eigenvalues	66
8.6 Remarks	68
REFERENCES	69

ABSTRACT

Methods of interpolation and quadrature have been used for over 300 years. Improvements in the techniques have been made by many, most notably by Gauss, whose technique applied to polynomials is referred to as Gaussian Quadrature. Stieltjes extended Gauss's method to certain non-polynomial functions as early as 1884. Conditions that guarantee the existence of quadrature formulas for certain collections of functions were studied by Tchebycheff, and his work was extended by others. Today, a class of functions which satisfies these conditions is called a Tchebycheff System. This thesis contains the definition of a Tchebycheff System, along with the theorems, proofs, and definitions necessary to guarantee the existence of quadrature formulas for such systems.

Solutions of discretely observable linear control systems are of particular interest, and observability with respect to a given output function is defined. The output function is written as a linear combination of a collection of orthonormal functions. Orthonormal functions are defined, and their properties are discussed.

The technique for evaluating the coefficients in the output function involves evaluating the definite integral of functions which can be shown to form a Tchebycheff system. Therefore, quadrature formulas for these integrals exist, and in many cases are known.

The technique given is useful in cases where the method of direct calculation is unstable. The condition number of a matrix is defined and shown to be an indication of the degree to which perturbations in data affect the accuracy of the solution. In special cases, the number of data points required for direct calculation is the same as the number required by the method presented in this thesis. But the method is shown to require more data points in other cases. A lower bound for the number of data points required is given.

CHAPTER I

INTRODUCTION

In the study of discretely observable linear control systems, a system of first order linear differential equations $\dot{x} = Ax$ is given, where A is an $n \times n$ constant matrix, and $x(t) = (x_1(t), x_2(t), \dots, x_n(t)) \in \mathbb{R}^n$ for t in an interval I . An output function y is a linear combination of the coordinates of x ; i. e., $y(t) = c_1 x_1(t) + c_2 x_2(t) + \dots + c_n x_n(t)$. A system is said to be observable if the initial value $x(t_0)$ is uniquely determined by the output function y . The system is said to be discretely observable if the values of $y(t)$ and some of its derivatives are known for a discrete set of values of t . (The data points required will be defined in Chapter V.)

The output function y typically has the form

$$y(t) = \sum_{i=1}^s r_i(t) e^{\lambda_i t}, \quad (1.1)$$

where for each i , r_i is a polynomial to be determined, and λ_i is a constant. The polynomials in the solution can be calculated directly if an adequate number of data points are given. But two problems arise if the dimension of the system is large. First, a direct calculation would require the inversion of a matrix of high order, and this might not be practical. Second, even if the inversion of the matrix is carried out, the accuracy of the solution depends upon the accuracy of the data being used. For some matrices, slight perturbations in the data can lead to large errors in the solution.

In this thesis, the solution y will be re-formulated using a new matrix known to be stable; that is, one which is minimally sensitive to perturbations in the data. Also, by using a collection of orthogonal functions, the constants in the solution can be expressed as integrals which can be calculated from the data points using quadrature methods, thus avoiding the necessity of actually inverting the matrix.

Many methods of approximating the value of a definite integral have long been known. Integration formulas of interpolatory type are those found by approximating an integrand with a polynomial and then integrating the polynomial. The approximating polynomial is one for which the values of the polynomial and the function are equal at a set of distinct points in the interval of integration. This method yields the

approximating formula

$$\int_I f(x) dx \approx \sum_{k=0}^n w_k f(x_k).$$

If the function f is a polynomial of degree less than or equal to n then the formula is exact.

Functions other than polynomials have the property that their integrals can be expressed exactly in this way. Suppose f is a function and t_1, t_2, \dots, t_n are points in an interval I . If

$$\int_I f(x) dx = \sum_{k=0}^n w_k f(x_k)$$

then $\sum_{k=0}^n w_k f(x_k)$ is said to be a quadrature formula. The points t_1, t_2, \dots, t_n are called nodes, and the constants w_1, w_2, \dots, w_n are called weights.

In general, there exist classes of functions for which a single quadrature formula is exact for each function in the class. Let $u_1, u_2, \dots, u_n, \dots$ be functions which are integrable over an interval I . If there exist points t_1, t_2, \dots, t_m in I and constants w_1, w_2, \dots, w_m such that

$$\int_I u_i(t) dt = \sum_{k=0}^m w_k u_i(t_k)$$

for each $i = 1, 2, \dots, n$ then the rule will be said to have degree of exactness n . If the functions u_i are the polynomials $1, t, t^2, t^3, \dots$, and t_1, t_2, \dots, t_n are distinct points in I , there exist weights w_1, w_2, \dots, w_n such that for all polynomials f of degree less than or equal to $n - 1$ the quadrature formula is exact. If the points t_1, t_2, \dots, t_n are chosen appropriately, the degree of exactness of the formula can be as great as $2n - 1$.

In the latter part of the 17th century, Newton approximated a function f by constructing the unique polynomial of degree less than or equal to $n - 1$ which passed through a set of n distinct points of the function. The polynomial was expressed in terms of divided differences, but the form which we find convenient today is the one devised in 1795 by Lagrange.

After learning of Newton's ideas, Roger Cotes approximated the integral of a function with the integral of the interpolating polynomial which agreed with the function at equally spaced nodes. He computed the weights w_1, w_2, \dots, w_n in the quadrature formula for all $n \leq 11$. The trapezoidal and Simpson rules are special cases.

Gauss raised the question of what the maximum degree of exactness would be if the nodes could be chosen arbitrarily. He used his theory of continued fractions associated with hypergeometric series to show in 1814 a way of choosing n nodes so that the quadrature formula is exact for all polynomials of degree less than or equal to $2n - 1$.

Jacobi showed in 1826 that if the nodes t_1, t_2, \dots, t_n can be chosen so that the node polynomial ω_n , given by

$$\omega_n(t) = (t - t_1)(t - t_2) \cdots (t - t_n)$$

is orthogonal to all polynomials of degree less than or equal to $k - 1$, then the quadrature rule has degree of exactness $n - 1 + k$. (A polynomial of degree -1 is taken to be identically zero.) Since ω_n cannot be orthogonal to itself, $k \leq n$, and the maximum degree of exactness possible is $2n - 1$.

It was not until the latter half of the nineteenth century that the work of Gauss and Jacobi was extended to weighted integrals of the form

$$\int_a^b f(t) d\sigma(t) = \int_a^b f(t) w(t) dt.$$

The technique of Christoffel (1877) depended upon the generation of orthogonal polynomials, and he showed that they satisfy a three-term recurrence relation.

In 1884, Stieltjes extended Gaussian quadrature to functions other than polynomials. In particular, he established formulas for the class of functions $u_k(t) = t^{\alpha_k}$, where $0 \leq \alpha_1 < \alpha_2 < \alpha_3 < \cdots$, on the interval $[0, 1]$.

The theory can be extended to a wide collection of classes of functions. In particular, quadrature formulas theoretically exist for any class of functions which forms what is now called a Tchebycheff system. (Such a system will be defined in Chapter 6.) Tchebycheff studied these systems in the latter half of the nineteenth century, and his work was extended by others, including M. G. Krein, S. Karlin, and L. S. Shapley. A comprehensive presentation of Tchebycheff systems is given in [28].

The output function (1.1) is the sum of functions of the form

$$u_{j,k}(t) = t^j \exp(\lambda_k t), \quad \text{for nonnegative integers } j \text{ and } k. \quad (1.2)$$

This class of functions forms a Tchebycheff system, and the method of computing the polynomials $r_i(t)$ to be presented in this thesis will require quadrature formulas

for integrals of functions of the form (1.2). The definitions and properties of orthogonal functions, Gaussian quadrature, stability of matrices, observability, Tchebycheff systems, and exponential interpolation will be discussed before discussing the application of these ideas to the determination of the output function (1.1).

CHAPTER II

ORTHOGONAL FUNCTIONS

In many applications, a collection of orthogonal functions can be used to demonstrate existence of solutions and to facilitate computations. Because of their importance in quadrature methods, a summary of some of their properties will be presented in this chapter. For a more detailed discussion of most of what follows, see [9].

Definition 2.1 *Let X be an inner product space with inner product of f and g denoted by $\langle f, g \rangle$. Two elements $f, g \in X$ are said to be orthogonal if $\langle f, g \rangle = 0$. A subset S of X is said to be an orthogonal set if for all $f, g \in S$ such that $f \neq g$, f and g are orthogonal.*

Definition 2.2 *Let S be an orthogonal subset of an inner product space X . If $\langle f, f \rangle = 1$ for all $f \in S$, then S is said to be orthonormal.*

A closed bounded interval $[a, b]$ and a real-valued function w determine an inner product space defined as follows.

Definition 2.3 *Let $X = C[a, b]$. Let w belong to X and $w(x) > 0$ on $[a, b]$. For each pair $f, g \in X$, define*

$$\langle f, g \rangle = \int_a^b f(t)g(t)w(t)dt. \quad (2.1)$$

Then $\langle f, g \rangle$ exists for all $f, g \in X$ and is called the inner product of f and g with respect to the weight function w .

The space X defined above is an inner product space, and a set $\{\phi_k\}$ of orthogonal polynomials with respect to a positive weight function w can always be found, where the degree of ϕ_k is equal to k . Inner product spaces also exist in case the interval of integration is infinite if the set X is restricted to integrable functions.

One method of generating a set of orthonormal polynomials is the Gram-Schmidt process. This process can be applied to any sequence of functions x_1, x_2, \dots , provided the set $\{x_1, x_2, \dots, x_k\}$ is linearly independent for each k . Assume $x_0(t) \equiv 1$ and $x_k(t) = t^k$ for each positive integer k . Define

$$\phi_0(t) = \frac{x_0(t)}{\langle x_0, x_0 \rangle^{1/2}}$$

$$\begin{aligned}
y_1(t) &= x_1(t) - \langle x_1, \phi_0 \rangle \phi_0(t) \\
\phi_1(t) &= \frac{y_1(t)}{\langle y_1, y_1 \rangle^{1/2}} \\
&\vdots \\
y_k(t) &= x_k(t) - \sum_{k=0}^{k-1} \langle x_k, \phi_{k-1} \rangle \phi_{k-1} \\
\phi_k(t) &= \frac{y_k(t)}{\langle y_k, y_k \rangle^{1/2}} \\
&\vdots
\end{aligned}$$

The set $\{\phi_0, \phi_1, \dots, \phi_k, \dots\}$ is orthonormal. For each k , the degree of ϕ_k is equal to k , and the leading coefficient of ϕ_k is positive. The following conditions also hold.

1. For each $\phi_k(t)$, there exist constants a_0, a_1, \dots, a_k such that

$$\phi_k(t) = a_0 x_0(t) + a_1 x_1(t) + \dots + a_k x_k(t).$$

2. For each x_k , there exist constants b_0, b_1, \dots, b_k such that

$$x_k(t) = b_0 \phi_0(t) + b_1 \phi_1(t) + \dots + b_k \phi_k(t).$$

This implies that ϕ_k is orthogonal to every polynomial of degree less than k . Suppose $\{\phi_k\}$ and $\{x_k\}$ are collections of polynomials for which condition (1) above holds. If $\{\phi_k\}$ is an orthonormal set of polynomials with positive leading coefficients, then it is precisely the set of polynomials which is generated from the collection $\{x_k\}$ by the Gram-Schmidt process.

Real orthonormal polynomials satisfy a recursion relation, and the following theorem provides a method of computing the polynomials if the constants in the recursion relation can be calculated.

Theorem 2.1 *Real orthonormal polynomials satisfy the following three term recursion relation.*

$$\phi_n(t) = (\alpha_n t + \beta_n) \phi_{n-1}(t) - \gamma_n \phi_{n-2}(t)$$

The following theorem gives a method of constructing an orthogonal set of monic polynomials. In this method, constants c_0, c_1, c_2, \dots can be defined by

$$\begin{aligned}
c_0 &= \int_a^b w(t) dt \\
c_n &= \int_a^b t^n w(t) dt, \quad n = 1, 2, \dots,
\end{aligned}$$

but the theorem is valid for more general sequences $\{c_n\}$, provided

$$s_n = \det \begin{pmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_1 & c_2 & \cdots & c_n \\ & & \cdots & \\ c_{n-1} & c_n & \cdots & c_{2n-2} \end{pmatrix} \neq 0 \quad \text{for } n = 1, 2, \dots \quad (2.2)$$

In general, let c^* be the linear functional on the space of real polynomials defined by

$$c^*(t^n) = c_n, \quad n = 0, 1, 2, \dots$$

An inner product on this space is defined by

$$\langle p, q \rangle = c^*(p(t)q(t)) \quad (2.3)$$

Theorem 2.2 *There exists a unique sequence of monic polynomials which form an orthogonal set with respect to the inner product [2.3]. The polynomials are given by the following, where s_n is defined for $n = 1, 2, \dots$ by [2.2]:*

$$q_n = \frac{1}{s_n} \det \begin{pmatrix} c_0 & c_1 & \cdots & c_n \\ c_1 & c_2 & \cdots & c_{n+1} \\ & & \cdots & \\ c_{n-1} & c_n & \cdots & c_{2n-1} \\ 1 & t & \cdots & t^n \end{pmatrix}.$$

The sequence of polynomials $\{q_n\}$ determined in the above theorem satisfies the three term recurrence relation

$$\begin{aligned} q_{-1}(t) &\equiv 0, \quad q_0(t) \equiv 1, \\ q_{n+1}(t) &= (t - \alpha_{n+1})q_n(t) - \beta_n^2 q_{n-1}(t) \\ n &= 0, 1, 2, \dots \end{aligned} \quad (2.4)$$

An algorithm for computing the constants α_n and β_n is given in [19]. These polynomials are called *Lanczos polynomials of the first kind* for $\{c_n\}$. *Lanczos polynomials of the second kind* for $\{c_n\}$, denoted by $\{p_n\}$, are generated if the initial conditions in [2.4] are altered as follows:

$$\begin{aligned} p_{-1} &\equiv -1, \quad p_0(t) \equiv 0, \\ p_{n+1}(t) &= (t - \alpha_{n+1})p_n(t) - \beta_n^2 p_{n-1}(t), \\ n &= 0, 1, 2, \dots \end{aligned} \quad (2.5)$$

The following classical examples of real orthogonal polynomials are generated by the inner product [2.1], where $[a, b] = [-1, 1]$.

1. The weight function $w(t) \equiv 1$ yields the Legendre polynomials.
2. The weight function $(1 - t^2)^{-1/2}$ produces Tschebysheff polynomials of the first kind.
3. The weight function $(1 - t^2)^{1/2}$ produces Tschebysheff polynomials of the second kind.
4. Jacobi polynomials are produced by the weight function $(1 - t)^\alpha(1 + t)^\beta$, where α and β are constants greater than -1 .

For some weight functions, the interval may be chosen to be infinite. For example,

1. generalized Laguerre polynomials result when the interval is $[0, \infty)$ and the weight function is $t^\alpha e^{-t}$, where α is a constant greater than -1 , and
2. Hermite polynomials occur for an interval $(-\infty, \infty)$ and a weight function e^{-t^2} .

An important property of the zeros of real orthogonal polynomials with respect to the inner product [2.1] is given in the following theorem.

Theorem 2.3 *The zeros of real orthogonal polynomials are real, distinct, and are located in the open interval (a, b) .*

The following definition of the kernel polynomial of an orthonormal system and the Christoffel-Darboux theorem lead to an interlacing property of the zeros of consecutive polynomials.

Definition 2.4 *Let $\{\phi_n\}$ be a system of real orthonormal polynomials. The function*

$$K_n(t, s) = \sum_{k=0}^n \phi_k(t)\phi_k(s)$$

is called the kernel polynomial of order n of the orthonormal system.

Theorem 2.4 (Christoffel-Darboux) *Let $\{\phi_n\}$ be a set of real orthonormal polynomials, where*

$$\phi_n(t) = a_{nn}t^n + a_{n(n-1)}t^{n-1} + \cdots + a_{n0}$$

for $n = 0, 1, 2, \dots$. Then

$$K_n(t, s) = \left(\frac{a_{nn}}{a_{(n+1)(n+1)}} \right) \left(\frac{\phi_{n+1}(t)\phi_n(s) - \phi_n(t)\phi_{n+1}(s)}{t-s} \right)$$

$$\text{and } \sum_{k=0}^n \{\phi_k(t)\}^2 = \left(\frac{a_{nn}}{a_{(n+1)(n+1)}} \right) (\phi'_{n+1}(t)\phi_n(t) - \phi'_n(t)\phi_{n+1}(t))$$

Using the Christoffel-Darboux theorem, the following can be proved.

Theorem 2.5 *Let $\{\phi_n\}$ be a collection of orthogonal polynomials where, for $n = 0, 1, 2, \dots$, the degree of ϕ_n is equal to n . Let k be a positive integer. By theorem [2.3], ϕ_{k+1} has $k+1$ real distinct zeros. If λ_1 and λ_2 are consecutive zeros of ϕ_{k+1} , then ϕ_k has a zero between λ_1 and λ_2 .*

CHAPTER III

GAUSSIAN QUADRATURE

The problem of approximating a definite integral has been studied extensively, and many methods have been developed. The method of quadrature takes advantage of the fact that the polynomials are dense in the set of all functions which are continuous on a given interval. If the value of a function f is known at points t_1, t_2, \dots, t_n in the interval of integration, a polynomial of degree $n - 1$ exists which agrees with the function f at each of these points. Such a polynomial is called an interpolating polynomial for f , and is said to interpolate f at the points t_1, t_2, \dots, t_n . The integral of this interpolating polynomial is then an approximation of the integral of the function f .

The computation of the integral of an interpolating polynomial for the function f at points t_1, t_2, \dots, t_n requires the determination of constants $\lambda_1, \lambda_2, \dots, \lambda_n$ as well as the values $f(t_1), f(t_2), \dots, f(t_n)$. The integral is then given by the *quadrature formula* $\sum_{k=1}^n \lambda_k f(t_k)$. If two points t_1 and t_2 are used, the interpolating polynomial is linear, and the quadrature formula is a special case of the trapezoidal rule. If three equally spaced points are used, the quadrature formula is a special case of Simpson's rule.

An advantage of the quadrature method is that the quadrature formula is exact for all functions which are polynomials of degree less than or equal to $n - 1$. If some of the points t_1, t_2, \dots, t_n are not predetermined, they can be chosen so that the quadrature formula will be exact for polynomials of higher degree.

In general, suppose u_1, u_2, \dots, u_n are continuous real-valued functions defined on a closed finite interval $[a, b]$ containing the points t_1, t_2, \dots, t_m . Let $\lambda_1, \lambda_2, \dots, \lambda_m$ be constants such that for each $k = 1, 2, \dots, n$,

$$\int_a^b u_k(t) dt = \sum_{j=1}^m \lambda_j u_k(t_j).$$

If the function f is a linear combination of the functions u_1, u_2, \dots, u_n , then

$$\int_a^b f(t) dt = \sum_{j=1}^m \lambda_j f(t_j).$$

The constants $\lambda_1, \lambda_2, \dots, \lambda_m$ are called *weights*, and the points t_1, t_2, \dots, t_m are called *nodes*. Necessary and sufficient conditions for the existence of such nodes and weights

for a given set of functions, along with the number of nodes and weights required, are given in [28] and will be discussed elsewhere in this paper.

Consider the particular case in which the functions to be integrated are polynomials. If the nodes are pre-determined, then n of them will be required for the quadrature formula to be exact for all polynomials of degree less than or equal to $n - 1$. If, however, all the nodes may be selected appropriately, then n nodes will give a quadrature formula exact for all polynomials of degree less than or equal to $2n - 1$.

Let n be a positive integer and \mathcal{P}_{n-1} the set of all polynomials of degree less than or equal to $n - 1$. Given n distinct points t_1, t_2, \dots, t_n in the interval $[a, b]$ and n values y_1, y_2, \dots, y_n , there exists a unique polynomial $p \in \mathcal{P}_{n-1}$ for which

$$p(t_k) = y_k \quad k = 1, 2, \dots, n.$$

The polynomial p can be represented by the Lagrange formula as follows:

$$p(t) = \sum_{k=1}^n y_k \ell_k(t),$$

where for each k , ℓ_k is the Lagrange polynomial defined by

$$\ell_k(t) = \frac{(t - t_1)(t - t_2) \cdots (t - t_{k-1})(t - t_{k+1}) \cdots (t - t_n)}{(t_k - t_1)(t_k - t_2) \cdots (t_k - t_{k-1})(t_k - t_{k+1}) \cdots (t_k - t_n)}. \quad (3.1)$$

Then

$$\int_a^b p(t) dt = \sum_{k=1}^n \lambda_k y_k$$

where for each $k = 1, 2, \dots, n$,

$$\lambda_k = \int_a^b \ell_k(t) dt.$$

Since the scalar values λ_k depend only upon the nodes t_1, t_2, \dots, t_n , then for each polynomial in \mathcal{P}_{n-1} , the integral can be evaluated if the value of the polynomial is known at each of these nodes.

Suppose the nodes t_1, t_2, \dots, t_n in the above method are not pre-determined. It was found by Gauss that they can be selected so that the quadrature formula is exact for all polynomials in \mathcal{P}_{2n-1} . In order to determine such nodes, a set of orthogonal polynomials is defined, and the nodes are roots of the polynomial of degree n .

By the theory of orthogonal polynomials, if $\{\phi_0, \phi_1, \dots, \phi_n\}$ is a set of real orthogonal polynomials over the interval $[a, b]$ such that for each $k = 0, 1, \dots, n$, the degree

of ϕ_k is equal to k , then the roots of each ϕ_k are real, distinct, and lie in the interval (a, b) . Further, if p is any polynomial of degree less than k , then ϕ_k is orthogonal to p . The following theorem validates the method of Gaussian quadrature.

Theorem 3.1 (Gauss-Jacobi) *Let w be a positive weight function and $\{\phi_0, \phi_1, \dots, \phi_n\}$ a set of orthogonal polynomials with respect to the inner product (2.3) such that for each $k = 0, 1, \dots, n$, the degree of ϕ_k is equal to k . Let t_1, t_2, \dots, t_n be the zeros of ϕ_n , where $a < t_1 < t_2 < \dots < t_n < b$. There exist positive constants $\lambda_1, \lambda_2, \dots, \lambda_n$ such that $\int_a^b p(t)w(t)dt = \sum_{k=1}^n \lambda_k p(t_k)$ whenever $p \in \mathcal{P}_{2n-1}$.*

Proof: Let p belong to \mathcal{P}_{2n-1} , and define $q \in \mathcal{P}_{n-1}$ by

$$q(t) = \sum_{k=1}^n p(t_k) \ell_k(t)$$

where $\ell_k(t)$ is the Lagrange polynomial (3.1). Since $\ell_k(t)$ satisfies

$$\ell_k(t_j) = \delta_{kj} = \begin{cases} 0 & \text{if } k \neq j \\ 1 & \text{if } k = j \end{cases}$$

then $q(t_k) = p(t_k)$ for each $k = 1, 2, \dots, n$. Therefore $p(t) - q(t)$ belongs to \mathcal{P}_{2n-1} and has zeros at t_1, t_2, \dots, t_n . Since these are precisely the zeros of the n th degree polynomial ϕ_n , we have

$$p(t) - q(t) = \phi_n(t) r_{n-1}(t)$$

where $r_{n-1} \in \mathcal{P}_{n-1}$. Since the degree of r_{n-1} is less than the degree of ϕ_n , then ϕ_n is orthogonal to r_{n-1} , and therefore

$$\begin{aligned} \int_a^b p(t)w(t)dt &= \int_a^b [q(t) + \phi_n(t)r_{n-1}(t)]w(t)dt \\ &= \int_a^b q(t)w(t)dt \\ &= \int_a^b \left[\sum_{k=1}^n p(t_k) \ell_k(t) \right] w(t)dt \\ &= \sum_{k=1}^n \left\{ p(t_k) \int_a^b \ell_k(t)w(t)dt \right\}. \end{aligned}$$

Define

$$\lambda_k = \int_a^b \ell_k(t)w(t)dt. \tag{3.2}$$

Then

$$\int_a^b p(t)w(t)dt = \sum_{k=1}^n \lambda_k p(t_k). \quad (3.3)$$

To see that each λ_k is positive, note that $(\ell_k(t))^2 \in \mathcal{P}_{2n-2}$, and therefore

$$\int_a^b (\ell_k(t))^2 w(t)dt = \sum_{j=1}^n \lambda_j (\ell_k(t_j))^2.$$

Also, $(\ell_k(t))^2$ has zeros at $t_1, t_2, \dots, t_{k-1}, t_{k+1}, \dots, t_n$, and $(\ell_k(t_k))^2 = 1$. So

$$\begin{aligned} \lambda_k &= \sum_{j=1}^n \lambda_j (\ell_k(t_j))^2 \\ &= \int_a^b (\ell_k(t))^2 w(t)dt > 0. \end{aligned}$$

Although (3.2) defines the weights in the Gauss formula (3.3), other methods exist for their determination. For references, see Gautschi's survey [13].

CHAPTER IV

STABILITY

Let A be a nonsingular matrix in $\mathbb{C}^{n \times n}$ and b an element of \mathbb{C}^n . Then the equation $Ax = b$ has a unique solution for x , where x belongs to \mathbb{C}^n . In general, the accuracy of the solution x is affected by the accuracy of A and b . The problem is considered to be *stable*, or *well-conditioned* if only small perturbations in the solution x result from small perturbations in A and b .

In order to quantify the magnitude of a perturbation, it is convenient to define a vector norm $\| \cdot \|$. If x is the solution of $Ax = b$, and $x + y$ is an approximate solution, a convenient measure of the magnitude of the error is the *relative perturbation of x* , defined by $\|y\|/\|x\|$.

Suppose the matrix A and the vector b are perturbed by F and k respectively. If $x + y$ is the solution of the perturbed system, then

$$(A + F)(x + y) = b + k.$$

An upper bound for the relative error depends not only on the magnitude of the perturbation of A and b , but also on the matrix A . The following theorem will be helpful in determining an upper bound. Details of the proof of this theorem and the following results are given in [31].

Theorem 4.1 *If $\| \cdot \|$ denotes any matrix norm for which $\|I\| = 1$, and if $\|M\| < 1$, then $(I + M)^{-1}$ exists and*

$$\|(I + M)^{-1}\| \leq \frac{1}{1 - \|M\|}.$$

The following definition of the condition number of a matrix will be useful in expressing an upper bound for the relative error.

Definition 4.1 *Let A be a nonsingular matrix. The condition number of A with respect to $\| \cdot \|$, denoted $C(A)$, is defined by*

$$C(A) = \|A\| \|A^{-1}\|$$

An upper bound for the relative error is given by the following theorem.

Theorem 4.2 *Let A and F belong to $\mathbb{C}^{n \times n}$, and let x, y, b , and K belong to \mathbb{C}^n . Assume that $(A + F)(x + y) = b + k$. Assume also that $\|A^{-1}\|\|F\| < 1$ and $\|I\| = 1$. Choose a vector norm which is compatible with the matrix norm. (That is, $\|Ax\| \leq \|A\|\|x\|$ for all x in \mathbb{C}^n and all A in $\mathbb{C}^{n \times n}$.) Then the relative error satisfies*

$$\|y\|/\|x\| \leq \frac{C(A)}{1 - (C(A)\|F\|/\|A\|)} \left(\frac{\|k\|}{\|b\|} + \frac{\|F\|}{\|A\|} \right)$$

As expected, the relative error in a solution x of $Ax = b$ will be small if k and F are “small” relative to b and A . But the condition number of A also determines the upper bound. If $C(A)$ is large, then it is possible that small perturbations in A and b could result in a rather large relative error in the solution x . Since

$$C(A) = \|A\|\|A^{-1}\| \geq \|A \cdot A^{-1}\| = \|I\| = 1,$$

then a well-conditioned system would have $C(A)$ close to 1.

If there is no perturbation of the matrix A , the above upper bound for the relative perturbation of x reduces to

$$\|y\|/\|x\| \leq C(A) \left(\frac{\|k\|}{\|b\|} \right)$$

In case the only perturbation is in the matrix A , then

$$\frac{\|y\|}{\|x\|} \leq \frac{C(A)}{\frac{\|A\|}{\|F\|} - C(A)}$$

Let $A = (a_{ij})$ belong to $\mathbb{C}^{n \times n}$ and define the matrix norm $\|\cdot\|_1$ by

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|.$$

Consider an $n \times n$ complex matrix generated by the functions $e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_n t}$, where $\lambda_1, \lambda_2, \dots, \lambda_n$ are complex constants such that $\operatorname{Re} \lambda_1 \leq \operatorname{Re} \lambda_2 \leq \dots \leq \operatorname{Re} \lambda_n$. For real numbers t_1, t_2, \dots, t_n satisfying $0 \leq t_1 < t_2 < \dots < t_n < \infty$, define the matrix

$$E_n = \begin{pmatrix} e^{\lambda_1 t_1} & e^{\lambda_2 t_1} & \dots & e^{\lambda_n t_1} \\ e^{\lambda_1 t_2} & e^{\lambda_2 t_2} & \dots & e^{\lambda_n t_2} \\ & & \dots & \\ e^{\lambda_1 t_n} & e^{\lambda_2 t_n} & \dots & e^{\lambda_n t_n} \end{pmatrix}$$

It is shown in [32] that if E_n is nonsingular, then the condition number of E_n is greater than or equal to n . If $\lambda_1, \lambda_2, \dots, \lambda_n$ are real and distinct, E_n will be nonsingular for any choice of (t_1, t_2, \dots, t_n) provided $0 \leq t_1 < t_2 < \dots < t_n < \infty$. Therefore, the existence of a choice for (t_1, t_2, \dots, t_n) which produces a matrix E_n with minimum condition number is assured. The paper also produces upper and lower bounds for $C(E_n)$ for some special cases.

CHAPTER V

OBSERVABILITY

Let A be a constant matrix in $\mathbb{C}^{n \times n}$, c a constant vector in \mathbb{R}^n , and I an interval. Let x be a function from I to \mathbb{R}^n . Consider the linear system

$$\dot{x} = Ax \quad y = c^T x \quad x(t_0) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad t_0 \in I \quad (5.1)$$

The function y is called the *output function*. In most of what follows, it will be assumed that $t_0 = 0$, since an appropriate translation of the variable transforms a system into this form. It will also be assumed that I has the form $[0, \infty)$ or $[0, b]$ for a real number $b > 0$.

Let $x_1, x_2 \in \mathbb{R}^n$ be values for $x(t_0)$ and the output functions y_1, y_2 solutions of (5.1) corresponding to the initial values x_1 and x_2 respectively. The system is said to be *observable* when $y_1 \equiv y_2$ if and only if $x_1 = x_2$. An observable system is said to be *discretely observable* with respect to points t_1, t_2, \dots, t_m in the interval I and nonnegative integers s_1, s_2, \dots, s_m satisfying $\sum_{k=1}^m s_k = n$ if whenever y is a solution of (5.1) satisfying

$$\begin{aligned} y(t_1) &= y(t_2) = \dots = y(t_m) = 0 \text{ and if } m < n \\ y'(t_k) &= y''(t_k) = \dots = y^{[s_k-1]}(t_k) = 0 \\ &\text{for each } k \text{ where } s_k > 1 \end{aligned} \quad (5.2)$$

then $y(t) \equiv 0$ on I .

If the system (5.1) is observable, then the constant vector c has no zero components. The solution for x is given by

$$x(t) = \exp(At)x(0).$$

The $n \times n$ matrix $\exp(At)$ has the form $\exp(At) = (q_{ij}(t))$, where each $q_{ij}(t)$ is given

by

$$q_{ij}(t) = \sum_{k=1}^s p_{ijk}(t) e^{\lambda_k t}$$

where $\{\lambda_1, \lambda_2, \dots, \lambda_s\}$ is the set of distinct eigenvalues of A , and each p_{ijk} is a polynomial of degree less than n . The output function y is given by

$$y(t) = c^T x(t) = c^T \exp(At) x(0) = \sum_{i,j=1}^n b_{ij} \cdot q_{ij}(t)$$

where each b_{ij} is a constant. Regrouping the terms gives $y(t) = \sum_{i=1}^s r_i(t) e^{\lambda_i t}$, where each r_i is a polynomial of degree less than n .

The eigenvalues of the matrix A determine the form of the solution y , and it will be assumed that A is in Jordan canonical form. This assumption leads to no loss of generality since every matrix is similar to a matrix J in Jordan canonical form, and the linearity of the dynamical system is preserved under a change of variables of the form $Z = Px$, where $A = P^{-1}JP$.

Suppose a set $\{t_i\}_1^n \subset I$ satisfies (5.2). Assume the eigenvalues of the matrix A are distinct and that

$$A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}.$$

In this case, the output function y is given by

$$y(t) = \sum_{i=1}^n b_i e^{\lambda_i t}$$

for constants b_1, b_2, \dots, b_n . Then $y(t) \equiv 0$ if and only if $b_i = 0$ for each $i = 1, 2, \dots, n$. This will hold if and only if the system

$$\begin{pmatrix} e^{\lambda_1 t_1} & e^{\lambda_2 t_1} & \dots & e^{\lambda_n t_1} \\ e^{\lambda_1 t_2} & e^{\lambda_2 t_2} & \dots & e^{\lambda_n t_2} \\ & & \dots & \\ e^{\lambda_1 t_n} & e^{\lambda_2 t_n} & \dots & e^{\lambda_n t_n} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

has only the zero solution for the vector $(b_1, b_2, \dots, b_n)^T$. Therefore the system will be discretely observable with respect to t_1, t_2, \dots, t_n if and only if the matrix

$$E = \begin{pmatrix} e^{\lambda_1 t_1} & e^{\lambda_2 t_1} & \dots & e^{\lambda_n t_1} \\ e^{\lambda_1 t_2} & e^{\lambda_2 t_2} & \dots & e^{\lambda_n t_2} \\ & & \dots & \\ e^{\lambda_1 t_n} & e^{\lambda_2 t_n} & \dots & e^{\lambda_n t_n} \end{pmatrix}$$

is nonsingular.

Suppose the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of A are all real. It is shown in [28] that the determinant of E is positive provided $\lambda_1 < \lambda_2 < \dots < \lambda_n$ and $t_1 < t_2 < \dots < t_n$. So, in this case, the system would be observable for any choice of n distinct points.

If some of the eigenvalues of A are not real, it is shown in [39] that the determinant of E is nonzero if t_1 is chosen so that $t_1 \neq 2m\pi/(\lambda_k - \lambda_j)$ for all integers m and all $k \neq j$, where $1 \leq k, j \leq n$, and for $t_k = kt_1$, where $k = 2, \dots, n$. The thesis also provides some examples of discretely observable systems in case A belongs to $\mathbb{C}^{2 \times 2}$ or $\mathbb{C}^{3 \times 3}$.

The question of discrete observability is related to an interpolation problem defined as follows.

Definition 5.1 *Let V be an n -dimensional linear space. Let L_1, L_2, \dots, L_n be given linear functionals defined on V , and let w_1, w_2, \dots, w_n be given constants. The set of ordered pairs $\{(L_k, w_k)\}_{k=1}^n$ defines an interpolation problem. If $x \in V$, and $L_k(x) = w_k$ for each $k = 1, 2, \dots, n$, then x is said to be a solution of the interpolation problem.*

Let $V = \{y \mid y \text{ is an output function of the system (5.1)}\}$. Then V is of dimension n if and only if the system is observable. For a proof of this and the following theorem, see [39].

Theorem 5.1 *Let $\{t_1, t_2, \dots, t_n\}$ be a set of discrete points in I , and define linear functionals L_1, L_2, \dots, L_n by $L_k(y) = y(t_k)$. Then the interpolation problem $\{(L_k, 0)\}_{k=1}^n$ has a unique solution if and only if the system (5.1) is discretely observable with respect to t_1, t_2, \dots, t_n .*

Consider a system with a nilpotent matrix A , where $A^n = 0$. In this case, the output function y can be shown to be a polynomial of degree less than or equal to

$n - 1$. The interpolation problem has a unique solution for any choice of points t_1, t_2, \dots, t_n . (See [8].) Therefore, the system is discretely observable with respect to these points.

Another condition which is equivalent to discrete observability is given by the following theorem in [8].

Theorem 5.2 *Let V be an n -dimensional linear space and V^* its dual space. Let L_1, L_2, \dots, L_n be elements of V^* . The interpolation problem defined by $\{(L_k, w_k)\}_{k=1}^n$ possesses a solution for arbitrary values w_1, w_2, \dots, w_n if and only if the functionals L_1, L_2, \dots, L_n are linearly independent in V^* . The solution to the interpolation problem is unique.*

Once it has been established that a system is discretely observable with respect to a set of points t_1, t_2, \dots, t_n , determination of the output function y can be made in a number of ways depending upon the characteristics of the matrix A . The choice of a method is influenced by the size of A and by its eigenvalues. For example, if the matrix A belongs to $\mathbb{R}^{2 \times 2}$ and the eigenvalues λ_1 and λ_2 are real and distinct, then y is given by

$$y(t) = b_1 e^{\lambda_1 t} + b_2 e^{\lambda_2 t}.$$

Given any two nonnegative values t_1 and t_2 , with $y(t_1) = w_1$ and $y(t_2) = w_2$, the constants b_1 and b_2 satisfy

$$\begin{pmatrix} e^{\lambda_1 t_1} & e^{\lambda_2 t_1} \\ e^{\lambda_1 t_2} & e^{\lambda_2 t_2} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

In this case, the inverse of the matrix

$$E = \begin{pmatrix} e^{\lambda_1 t_1} & e^{\lambda_2 t_1} \\ e^{\lambda_1 t_2} & e^{\lambda_2 t_2} \end{pmatrix}$$

can be easily computed and the coefficients b_1 and b_2 determined with accuracy. But if the matrix E is large, computing the inverse is not practical. Also, as discussed in Chapter IV, it might be poorly conditioned. In that case the accuracy of the solution might be poor regardless of the method used to solve the above equation.

It is shown in [8] that if $W = C[a, b]$ and t_1, t_2, \dots, t_n are distinct points in $[a, b]$, then the linear functionals L_1, L_2, \dots, L_n defined by $L_k(f) = f(t_k)$ are independent

in $C[a, b]$. It is also shown that if \mathbb{X} is an n -dimensional linear space, then the dual space \mathbb{X}^* is also n -dimensional. These two theorems imply the following.

Theorem 5.3 *Let $V = \{y \mid y \text{ is an output function of the system (5.1)}\}$, and let t_1, t_2, \dots, t_n be discrete points in $[a, b]$. Let L be any linear functional in V^* . Define L_1, L_2, \dots, L_n in V^* by $L_k(y) = y(t_k)$. Then there exist constants b_1, b_2, \dots, b_n such that*

$$\begin{aligned} L(y) &= b_1 L_1(y) + b_2 L_2(y) + \dots + b_n L_n(y) \\ &= b_1 y(t_1) + b_2 y(t_2) + \dots + b_n y(t_n) \end{aligned}$$

The above theorem suggests the applicability of quadrature methods when the linear functional L is defined by

$$L(y) = \int_a^b y(t) dt.$$

This will be discussed further in Chapter VII.

CHAPTER VI

TCHEBYCHEFF SYSTEMS

In the method of Gaussian quadrature, orthogonal polynomials are used to determine nodes, t_1, t_2, \dots, t_n , and weights, $\lambda_1, \lambda_2, \dots, \lambda_n$, such that

$$\int_a^b p(t)w(t)dt = \sum_{k=1}^n \lambda_k p(t_k)w(t_k)$$

for all polynomials p of degree less than or equal to $2n - 1$. Theoretically, nodes and weights exist which make

$$\int_a^b f(t)w(t)dt = \sum_{k=1}^n \lambda_k f(t_k)w(t_k)$$

for other collections of functions provided they satisfy certain conditions. In this chapter an introduction to Tchebycheff systems will be presented, and it will be shown that quadrature formulas are possible for functions which form such a system.

Definition 6.1 Let u_0, u_1, \dots, u_n denote continuous real-valued functions defined on a closed finite interval $[a, b]$. The collection of these functions will be called a Tchebycheff system over $[a, b]$, abbreviated T-system, provided the determinants

$$U \begin{pmatrix} 0, & 1, & \dots, & n \\ t_0, & t_1, & \dots, & t_n \end{pmatrix} = \begin{vmatrix} u_0(t_0) & u_0(t_1) & \dots & u_0(t_n) \\ u_1(t_0) & u_1(t_1) & \dots & u_1(t_n) \\ \vdots & \vdots & & \vdots \\ u_n(t_0) & u_n(t_1) & \dots & u_n(t_n) \end{vmatrix} \quad (6.1)$$

are strictly positive whenever $a \leq t_0 < t_1 < \dots < t_n \leq b$. The collection of functions u_0, u_1, \dots, u_n will be referred to as a complete Tchebycheff system, abbreviated CT-system, if $\{u_0, u_1, \dots, u_r\}$ is a T-system for each $r = 0, 1, \dots, n$.

An example of a CT-system is the collection $\{u_i\}_0^n$ of functions defined by $u_0(t) \equiv 1$, and $u_i(t) = t^i$ for i a positive integer. For any choice t_0, t_1, \dots, t_n , where $t_0 < t_1 < \dots < t_n$, the determinant

$$U \begin{pmatrix} 0, & 1, & \dots, & n \\ t_0, & t_1, & \dots, & t_n \end{pmatrix} = \begin{vmatrix} 1 & 1 & \dots & 1 \\ t_0 & t_1 & \dots & t_n \\ & & \dots & \\ t_0^n & t_1^n & \dots & t_n^n \end{vmatrix}$$

is the Vandermonde determinant and has the value

$$\prod_{0 \leq i < j \leq n} (t_j - t_i) > 0.$$

The following theorem guarantees that the set $\{\exp(\alpha_i t)\}_0^n$ is a T-system on any closed interval if each α_i is real and $\alpha_0 < \alpha_1 < \cdots < \alpha_n$.

Theorem 6.1 *Let $\{x_i\}_0^n$ and $\{y_i\}_0^n$ be sets of real numbers where $x_0 < x_1 < \cdots < x_n$ and $y_0 < y_1 < \cdots < y_n$. Let*

$$E_n = \begin{pmatrix} \exp(x_0 y_0) & \exp(x_0 y_1) & \cdots & \exp(x_0 y_n) \\ \exp(x_1 y_0) & \exp(x_1 y_1) & \cdots & \exp(x_1 y_n) \\ & & \cdots & \\ \exp(x_n y_0) & \exp(x_n y_1) & \cdots & \exp(x_n y_n) \end{pmatrix} \quad (6.2)$$

The determinant of E_n is positive.

Proof: Let u_n be any function of the form

$$u_n(y) = \sum_{i=0}^n a_i e^{x_i y} \quad (6.3)$$

where $x_i \in \mathbb{R}$ and $a_i \in \mathbb{R}$ for each $i = 0, 1, \dots, n$, $x_0 < x_1 < \cdots < x_n$, and $\sum_{i=0}^n a_i^2 > 0$.

It will be shown by induction that $u_n(y)$ has at most n distinct real zeros. First, let $n = 0$. For any real numbers $a_0 \neq 0$ and x_0 , the function u_0 defined by $u_0 = a_0 e^{x_0 y}$ has no zero. Suppose any function u_k of the form (6.3) has at most k distinct real zeros. Let x_0, x_1, \dots, x_{k+1} and a_0, a_1, \dots, a_{k+1} be real numbers with $x_{k+1} > x_k$ and $a_{k+1} \neq 0$. Define

$$u_{k+1}(y) = \sum_{k=0}^{k+1} a_i e^{x_i y}.$$

Suppose u_{k+1} has $k + 2$ zeros. Then $u_{k+1}(y) \exp(-x_{k+1} y)$ has $k + 2$ zeros, and by Rolle's theorem, $\frac{d}{dy} [u_{k+1}(y) \exp(-x_{k+1} y)]$ must have $k + 1$ zeros. But

$$\begin{aligned} \frac{d}{dy} u_{k+1}(y) \exp(-x_{k+1} y) &= \frac{d}{dy} \sum_{i=0}^{k+1} a_i \exp[(x_i - x_{k+1})y] \\ &= \frac{d}{dy} \left[a_{k+1} + \sum_{i=0}^k a_i \exp[(x_i - x_{k+1})y] \right] \\ &= \sum_{i=0}^k a_i (x_i - x_{k+1}) \exp[(x_i - x_{k+1})y], \end{aligned}$$

which has at most k distinct real zeros by assumption.

Next, induction can be used to show that $\det(E_n) \neq 0$. Given real numbers x_0 and y_0 , define $E_0 = (\exp(x_0 y_0))$. Then $\det(E_0) > 0$. Suppose $\det(E_{k-1}) \neq 0$ for all sets of real numbers $\{s_i\}_0^{k-1}$ and $\{t_i\}_0^{k-1}$, where $s_0 < s_1 < \dots < s_{k-1}$ and $t_0 < t_1 < \dots < t_{k-1}$. Given sets $\{x_i\}_0^k$ and $\{y_i\}_0^k$ satisfying $x_0 < x_1 < \dots < x_k$ and $y_0 < y_1 < \dots < y_k$, suppose $\det(E_k) = 0$. Define u_k by

$$u_k(y) = \det \begin{pmatrix} e^{x_0 y_0} & e^{x_0 y_1} & \dots & e^{x_0 y_{k-1}} & e^{x_0 y} \\ e^{x_1 y_0} & e^{x_1 y_1} & \dots & e^{x_1 y_{k-1}} & e^{x_1 y} \\ & & \dots & & \\ e^{x_k y_0} & e^{x_k y_1} & \dots & e^{x_k y_{k-1}} & e^{x_k y} \end{pmatrix}$$

Then u_k has $k+1$ distinct real zeros y_0, y_1, \dots, y_k . Expanding by the last column gives $u_k(y) = \sum_{i=0}^k a_i e^{x_i y}$. By assumption, each $a_i \neq 0$, so u_k can have at most k distinct real zeros. Therefore, $\det(E_k)$ cannot be zero.

Now that we have $\det(E_n) \neq 0$, induction can be used to show that it must be positive. First, $E_0 > 0$ for all real numbers x_0 and y_0 . Next, suppose $\det(E_{k-1}) > 0$ for all matrices of order k of the form (6.2) where $x_0 < x_1 < \dots < x_{k-1}$ and $y_0 < y_1 < \dots < y_{k-1}$. Let x_k be a real number and $x_k > x_{k-1}$. For any real number y , the determinant

$$\det(E_k(y)) = \begin{vmatrix} \exp(x_0 y_0) & \exp(x_0 y_1) & \dots & \exp(x_0 y_{k-1}) & \exp(x_0 y) \\ \exp(x_1 y_0) & \exp(x_1 y_1) & \dots & \exp(x_1 y_{k-1}) & \exp(x_1 y) \\ & & \dots & & \\ \exp(x_k y_0) & \exp(x_k y_1) & \dots & \exp(x_k y_{k-1}) & \exp(x_k y) \end{vmatrix}$$

can be evaluated by expanding by minors of the last column to get

$$\det(E_k(y)) = \sum_{i=0}^k a_i \exp(x_i y),$$

where the coefficient a_k is the determinant of a matrix of order k of the form (6.2) and therefore positive by assumption. For all $y > y_{k-1}$, $\det(E_k(y)) \neq 0$. So, on the interval (y_{k-1}, ∞) , either $\det(E_k(y)) > 0$ or $\det(E_k(y)) < 0$. Suppose $\det(E_k(y)) < 0$ on (y_{k-1}, ∞) . Then

$$\exp(-x_k y) \det(E_k(y)) = a_k + \sum_{i=0}^{k-1} a_i \exp[(x_i - x_k)y] < 0 \quad \text{for all } y > y_{k-1}.$$

But this is impossible since

$$\lim_{y \rightarrow \infty} \left(a_k + \sum_{i=0}^{k-1} a_i \exp[(x_i - x_k)y] \right) = a_k > 0$$

So $\det(E_k(y)) > 0$ for all $y > y_{k-1}$. \square

Other examples of T-systems can be constructed by noting that if $\{u_i\}_0^n$ is a T-system on $[a, b]$, the following hold.

1. If the function r is continuous and positive on $[a, b]$, then $\{ru_i\}_0^n$ is a T-system on $[a, b]$.
2. If the function r is continuous and increasing on $[c, d]$ and the range of r is $[a, b]$, then $\{u_i \circ r\}_0^n$ is a T-system on $[c, d]$.

Since $\{\exp^{\alpha_i t}\}_0^n$ is a T-system for $\alpha_0 < \alpha_1 < \dots < \alpha_n$ and $\ln t$ is continuous and increasing, (2) implies that $\{t^{\alpha_i}\}_0^n$ is a T-system on any closed subinterval of $(0, \infty)$.

The next definition generalizes the concept of a polynomial.

Definition 6.2 Let $\{u_i\}_0^n$ be a set of real-valued functions defined on the interval $[a, b]$. A function of the form $u = \sum_{i=0}^n a_i u_i$, where each a_i is real, is called a u -polynomial. A u -polynomial is said to be nontrivial if $\sum_{i=0}^n a_i^2 > 0$.

Example. A Dirichlet polynomial $\sum_{i=0}^n a_i e^{\alpha_i t}$ is a u -polynomial in the T-system $\{e^{\alpha_i t}\}_0^n$.

If $\{u_i\}_0^n$ is a T-system, the functions u_0, u_1, \dots, u_n are linearly independent, and therefore any u -polynomial, $u = \sum_{i=0}^n a_i u_i$, is uniquely determined by the coefficient vector (a_0, a_1, \dots, a_n) . Furthermore, a u -polynomial is determined by specifying its values at $n+1$ distinct points. In particular, any nontrivial u -polynomial has at most n distinct zeros.

Definition 6.3 The number of distinct zeros on an interval $[a, b]$ of a continuous function f is denoted by $Z(f)$.

Example. Let $u_0(t) \equiv 1$, and $u_i(t) = t^i$ for $i = 1, \dots, n$. As already mentioned, $\{u_i\}_0^n$ is a T-system on an interval $[a, b]$ and there exist real orthogonal polynomials $\phi_0, \phi_1, \dots, \phi_n$ such that for $i = 0, 1, \dots, n$, the degree of ϕ_i is equal to i . For each i , ϕ_i has exactly i real distinct zeros, all in the interval (a, b) . So $Z(\phi_i) = i$.

We know that if $\{u_i\}_0^n$ is a T-system, then $Z(u) \leq n$ for any u -polynomial. The following theorem gives a sufficient condition for a set of functions to be a T-system, which depends on knowledge of the zeros of all u -polynomials.

Theorem 6.2 *If $\{u_i\}_0^n$ is a T-system, then $Z(u) \leq n$ for every nontrivial u -polynomial u . Conversely, if a system $\{u_i\}_0^n$ of continuous functions on $[a, b]$ satisfies $Z(u) \leq n$ for every nontrivial u -polynomial u , then $\{u_i\}_0^n$ is a T-system except possibly for the sign of one of the functions.*

Proof: Let $\{u_i\}_0^n$ be a system of continuous functions on $[a, b]$ and $Z(u) \leq n$ for any nontrivial u -polynomial u . Define the function P , where

$$P : [a, b]^{n+1} \subset \mathbb{R}^{n+1} \rightarrow \mathbb{R} \text{ and}$$

$$P(t_0, t_1, \dots, t_n) = \begin{vmatrix} u_0(t_0) & u_0(t_1) & \cdots & u_0(t_n) \\ u_1(t_0) & u_1(t_1) & \cdots & u_1(t_n) \\ \vdots & \vdots & & \vdots \\ u_n(t_0) & u_n(t_1) & \cdots & u_n(t_n) \end{vmatrix}$$

Then P is continuous on the region $[a, b]^{n+1}$. If there exists a vector $(t_0, t_1, \dots, t_n)^T \in [a, b]^{n+1}$ such that $P(t_0, t_1, \dots, t_n) = 0$, then there would be a nonzero solution of

$$\begin{pmatrix} u_0(t_0) & u_0(t_1) & \cdots & u_0(t_n) \\ u_1(t_0) & u_1(t_1) & \cdots & u_1(t_n) \\ \vdots & \vdots & & \vdots \\ u_n(t_0) & u_n(t_1) & \cdots & u_n(t_n) \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = 0$$

For this nonzero solution, define

$$u(t) = c_0 u_0(t) + c_1 u_1(t) + \cdots + c_n u_n(t).$$

Then u is a u -polynomial with $n+1$ zeros, t_0, t_1, \dots, t_n . Since $Z(u) \leq n$ for every nontrivial u -polynomial, $P(t_0, t_1, \dots, t_n)$ must be nonzero for every choice $(t_0, t_1, \dots, t_n)^T \in [a, b]^{n+1}$. By the intermediate value theorem, $P(t_0, t_1, \dots, t_n)$ must have a fixed sign in $[a, b]^{n+1}$. \square

Corollary 6.1 *If $\{u_i\}_0^n$ is a CT-system, then for each $k = 0, 1, \dots, n$, $Z(u) \leq k$, whenever $u = \sum_{i=0}^k a_i u_i$, each a_i is real, and $\sum_{i=0}^k a_i^2 > 0$.*

An important property of a polynomial is the multiplicity of each of its zeros. If the multiplicity of a real zero is even, then the polynomial does not change signs at that point. The following definition categorizes the zeros of continuous functions according to whether or not the function changes signs at that zero.

Definition 6.4 For any continuous function f on $[a, b]$, an isolated zero $t_0 \in (a, b)$ of f is called a nonnodal zero provided f does not change sign at t_0 . All other zeros including zeros at the end points a and b are called nodal zeros.

Definition 6.5 The number of zeros on $[a, b]$ of a continuous function f , where nodal zeros are counted once and nonnodal zeros twice, is denoted by $\tilde{Z}(f)$.

Theorem 6.3 If $\{u_i\}_0^n$ is a T -system, then $\tilde{Z}(u) \leq n$ for every nontrivial u -polynomial u . Conversely, if $\{u_i\}_0^n$ is a system of continuous functions on $[a, b]$, and $\tilde{Z}(u) \leq n$ for every nontrivial u -polynomial u , then $\{u_i\}_0^n$ is a T -system except possibly for the sign of one of the functions.

Proof: The second part of the theorem follows from Theorem (6.2) since $Z(u) \leq \tilde{Z}(u)$. To prove the first part, assume $\tilde{Z}(u) \geq n + 1$ for some nontrivial u -polynomial u . If u has no nonnodal zero, then $\tilde{Z}(u) = Z(u)$, and Theorem (6.2) would be contradicted.

Assume u has at least one nonnodal zero. Let t_1, t_2, \dots, t_k be distinct real zeros of u with $t_1 < t_2 < \dots < t_k$. For each nonnodal zero t_i , choose $\epsilon_i > 0$ so that if $i \neq k$, then $t_i < t_i + \epsilon_i < t_{i+1}$, and if $i = k$, then $t_i < t_i + \epsilon_i < b$. If t_j is the first nonnodal zero (i.e., $t_j \leq t_i$ for all nonnodal zeros t_i), then choose $\epsilon_j > 0$ so that if $j \neq 1$, then $t_{j-1} < t_j - \epsilon_j$, and if $j = 1$, then $a < t_j - \epsilon_j$. Define the set $S = \{s | s \text{ is a zero of } u, \text{ or } s = t_i \pm \epsilon_i \text{ for a nonnodal zero } t_i \text{ of } u\}$. This set is finite and contains at least $n + 2$ elements. Label the elements of $S = \{s_0, s_1, \dots, s_M\}$ so that $s_i \leq s_k$ whenever $i \leq k$. Consider the points s_0, s_1, \dots, s_{n+1} . Then $u(s_i) \geq 0$ for i odd and $u(s_i) \leq 0$ for i even, or vice-versa. Also,

$$\begin{vmatrix} u(s_0) & u(s_1) & \cdots & u(s_{n+1}) \\ u_0(s_0) & u_0(s_1) & \cdots & u_0(s_{n+1}) \\ \vdots & \vdots & \ddots & \vdots \\ u_n(s_0) & u_n(s_1) & \cdots & u_n(s_{n+1}) \end{vmatrix} = 0$$

because the first row is a linear combination of the following rows. If the above determinant is expanded by the first row, we have

$$\sum_{i=0}^{n+1} a_i u(s_i) = 0.$$

Since $\{u_i\}_0^n$ is a T-system, each minor is positive, and therefore the a_i alternate in sign, as do the $u(s_i)$. Either $a_i u(s_i) \geq 0$ for all i or $a_i u(s_i) \leq 0$ for all i , so $a_i u(s_i)$ must be zero for all i . Since each $a_i \neq 0$, s_i must be a zero of u for $i = 0, 1, \dots, n+1$. But by theorem (6.2), $Z(u) \leq n$, so the assumption that $\tilde{Z}(u) \geq n+1$ is false. \square

In the proofs of some of the following theorems, the existence of a nonnegative u -polynomial which vanishes only at prescribed points will be assumed. The following theorem due to Krein guarantees the existence of such a function. Before stating the theorem, the following definition of a weight will be needed.

Definition 6.6 Let $T = \{t_1, \dots, t_k\}$ be an increasing set of distinct points in $[a, b]$. To each $t_i \in T$ a weight $\omega(t_i)$ is defined by

$$\omega(t_i) = \begin{cases} 2 & t_i \in (a, b) \\ 1 & t_i = a \text{ or } b \end{cases}$$

Theorem 6.4 (Krein) If $\{u_i\}_0^n$ is a T-system on $[a, b]$ and $\sum_{i=1}^k \omega(t_i) \leq n$, then there exists a nontrivial, nonnegative u -polynomial u vanishing precisely at the points of $T = \{t_1, \dots, t_k\}$. The only exception is that if n is even, and exactly one of the end points a or b is in T then $u(t)$ may vanish at the other end point as well.

Proof: First let $n = 2m + 1$ for m a positive integer. Suppose $a < t_1 < t_2 < \dots < t_k < b$. Select points $t_1', t_2', \dots, t_{m-k}'$ so that $t_k < t_1' < t_2' < \dots < t_{m-k}' < b$. Let the set $\{s_i\}_1^{2m+1}$ consist of the points

$$a, t_1, t_1 + \epsilon, t_2, t_2 + \epsilon, \dots, t_k, t_k + \epsilon, t_1', t_1' + \epsilon, \dots, t_{m-k}', t_{m-k}' + \epsilon \quad (6.4)$$

where $\epsilon > 0$ is chosen so that the above sequence is increasing and contained in $[a, b]$. Define the polynomial

$$u_\epsilon(t) = u \begin{pmatrix} 0, & 1, & \dots, & 2m, & 2m+1 \\ s_1, & s_2, & \dots, & s_{2m+1}, & t \end{pmatrix}$$

(6.5)

$$= \begin{vmatrix} u_0(s_1) & \cdots & u_0(s_{2m+1}) & u_0(t) \\ \vdots & & & \\ u_{2m+1}(s_1) & \cdots & u_{2m+1}(s_{2m+1}) & u_{2m+1}(t) \end{vmatrix}$$

Since $\{u_i\}_0^n$ is a T -system, $u_\epsilon(t)$ vanishes precisely on the set $\{s_i\}_1^{2m+1}$, and each zero is nodal. In particular,

$$\begin{aligned} u_\epsilon(t) > 0 & \quad \text{if} \quad s_{2i-1} < t < s_{2i} \quad (i = 1, 2, \dots, m) \\ & \quad \text{or} \quad s_{2m+1} < t \leq b. \end{aligned}$$

Expanding (6.5) by the last column gives

$$u_\epsilon(t) = \sum_{i=0}^n a_i(\epsilon) u_i(t)$$

where $K(\epsilon) = \sum_{i=0}^n [a_i(\epsilon)]^2$ is positive. Define the function

$$\begin{aligned} v(\epsilon, t) &= \frac{1}{\sqrt{K(\epsilon)}} \sum_{i=0}^n a_i(\epsilon) u_i(t) \\ &= \sum_{i=0}^n b_i(\epsilon) u_i(t), \quad \text{where } b_i(t) = \frac{a_i(\epsilon)}{\sqrt{K(\epsilon)}} \end{aligned}$$

Then $\sum_{i=0}^n [b_i(\epsilon)]^2 = 1$ for each positive ϵ suitably close to zero.

Select a sequence $\{\epsilon_j\}$ of small positive numbers so that $\epsilon_j \rightarrow 0$. Since $|b_i(\epsilon_j)| \leq 1$ for $i = 0, 1, \dots, n$ and j a positive integer, then there exists a subsequence $\{\epsilon_{j_k}\}$ such that $\{b_i(\epsilon_{j_k})\}$ converges for each $i = 0, 1, \dots, n$. Let a_0, a_1, \dots, a_n be limits of the subsequences, where

$$\{b_i(\epsilon_{j_k})\} \rightarrow a_i$$

and define

$$\tilde{u}(t) = \sum_{i=0}^n a_i u_i(t).$$

Then $\sum_{i=0}^n a_i^2 = 1$, and \tilde{u} vanishes at the points $a, t_1, t_2, \dots, t_k, t_1', t_2', \dots, t_{m-k}'$. All points except a are nonnodal zeros, so $\tilde{Z}(\tilde{u}) \geq 2m + 1$. By theorem (6.3), $\tilde{Z}(\tilde{u}) \leq 2m + 1$, so these are all the zeros of \tilde{u} .

Next, construct in a similar manner a nonnegative u -polynomial u^* that vanishes precisely at points $t_1, t_2, \dots, t_k, t_1'', t_2'', \dots, t_{m-k}'', b$, where the points in $\{t_j''\}_1^{m-k}$ are chosen so that $\{t_j'\}_1^{m-k} \cap \{t_j''\}_1^{m-k} = \emptyset$. Then the u -polynomial $u(t) = \tilde{u}(t) + u^*(t)$ is nonnegative and vanishes precisely at points t_1, \dots, t_k .

In case $\{t_1, \dots, t_k\}$ is such that either $t_1 = a$ or $t_k = b$, but not both, then the desired u -polynomial can be constructed in a manner similar to the above. If $t_1 = a$ and $t_k = b$, select points $t_1', t_2', \dots, t_{m-k+1}'$ satisfying $t_{k-1} < t_j' < t_k$ for $j = 1, \dots, m - k + 1$. As before, choose $\epsilon > 0$ small enough that the sequence $\{s_i(\epsilon)\}_1^{2m+1}$, consisting of points

$$a, t_1 + \epsilon, t_2, t_2 + \epsilon, \dots, t_{k-1}, t_{k-1} + \epsilon, t_1', t_1' + \epsilon, \dots, t_{m-k+1}', t_{m-k+1}' + \epsilon, b,$$

is increasing. Define $u_\epsilon(t)$ as before in (6.5). Letting $\epsilon \rightarrow 0$ as above, a nonnegative polynomial \tilde{u} is determined which vanishes at the points $t_1, \dots, t_{k-1}, t_1', \dots, t_{m-k+1}', t_k$. Since $t_1 = a$ and $t_k = b$, $\tilde{Z}(\tilde{u}) = 2m$. Since, $\tilde{Z}(\tilde{u}) \leq 2m + 1$ by Theorem 6.3, \tilde{u} cannot vanish elsewhere.

Next construct as before a nonnegative u -polynomial u^* which vanishes precisely at the points $t_1, \dots, t_{k-1}, t_1'', \dots, t_{m-k+1}'', t_k$, where $\{t_j'\}_1^{m-k+1} \cap \{t_j''\}_1^{m-k+1} = \emptyset$. Then $u(t) = \tilde{u}(t) + u^*(t)$ is a nonnegative u -polynomial which vanishes precisely at the points t_1, \dots, t_k .

In case n is even, an argument similar to the above produces a nonnegative u -polynomial u vanishing at precisely the points $T = \{t_1, \dots, t_k\}$ unless exactly one of the end points is in T . In that case the proof leaves open the possibility that u might also vanish at the other end point as well. \square

Theorem 6.5 *Let $\{u_i\}_0^n$ be a T -system. Let $T = \{t_1, t_2, \dots, t_k\}$ be a set of points in $[a, b]$, and let $\omega : T \rightarrow \{1, 2\}$ be such that $\omega(t_i) = 1$ if $t_i = a$ or $t_i = b$. Suppose $\sum_{i=1}^k \omega(t_i) \leq n$. Then there exists a u -polynomial u such that $u(t) \neq 0$ for $t \in (a, b) - T$ and such that t_i is a nodal zero if $\omega(t_i) = 1$ and a nonnodal zero if $\omega(t_i) = 2$. If n is odd, the polynomial vanishes precisely on the set T .*

In attempting to extend the method of Gaussian quadrature to functions other than polynomials, it will be helpful to define the *moment space* of a T -system. Its characterization depends upon a theorem of Carathéodory, the proof of which depends on the following theorem. (See [40].)

Theorem 6.6 *Let $A \subset \mathbb{R}^n$. The convex hull of A is the set of all finite convex combinations of elements of A . (A convex combination of points u_1, u_2, \dots, u_p in A is a linear combination of the form $\sum_{i=1}^p t_i u_i$, where $t_i \geq 0$ for all $i = 1, 2, \dots, p$, and $\sum_{i=1}^p t_i = 1$.)*

Proof: Let B be the set of all convex combinations of elements of A . Denote the convex hull of A by $\mathcal{C}(A)$. Since $\mathcal{C}(A)$ is a convex set containing A , it contains all finite convex combinations of elements of A . So $B \subset \mathcal{C}(A)$.

Suppose x and y are elements of B . Then x and y have representations of the form

$$x = \sum_{i=1}^n \lambda_i x_i \quad \text{and} \quad y = \sum_{i=1}^m \mu_i y_i$$

where $\lambda_i \geq 0$ for all $i = 1, \dots, n$, $\mu_i \geq 0$ for all $i = 1, \dots, m$, $\sum_{i=1}^n \lambda_i = 1$, and $\sum_{i=1}^m \mu_i = 1$. Let $t \in [0, 1]$. Then $(1 - t) \in [0, 1]$ and

$$tx + (1 - t)y = \sum_{i=1}^n (t\lambda_i)x_i + \sum_{i=1}^m [(1 - t)\mu_i]y_i$$

is a convex combination of elements of A . So B is a convex set, and B contains A . This gives $\mathcal{C}(A) \subset B$ and therefore $\mathcal{C}(A) = B$. \square

The proof of Carathéodory's theorem is given in [40], and it refers to the following theorem which is easily verified and therefore stated without proof.

Theorem 6.7 *Let x_1, \dots, x_p be points in V , where V is a linear space over \mathbb{R} . The following statements are equivalent:*

1. *For any j , the vectors $x_i - x_j, i \neq j$, are linearly independent.*
2. *If $\alpha_1, \dots, \alpha_p$ are real numbers such that*

$$\sum_{i=1}^p \alpha_i x_i = 0 \quad \text{and} \quad \sum_{i=1}^p \alpha_i = 0$$

then $\alpha_1 = \alpha_2 = \dots = \alpha_p = 0$.

Theorem 6.8 (Carathéodory) *Every point belonging to the convex hull of a given set A in \mathbb{R}^n can be represented as a convex combination involving at most $n + 1$ points of A .*

Proof: Let x belong to the convex hull of A . Then by Theorem 6.6, there exist $x_1, \dots, x_p \in A$ and $\lambda_1, \dots, \lambda_p \in \mathbb{R}$ such that $\lambda_i \geq 0$ for $i = 1, \dots, p$, $\sum_{i=1}^p \lambda_i = 1$, and

$$x = \sum_{i=1}^p \lambda_i x_i.$$

Suppose there exist $\mu_1, \dots, \mu_p \in \mathbb{R}$, not all zero, such that

$$\sum_{i=1}^p \mu_i x_i = 0 \text{ and } \sum_{i=1}^p \mu_i = 0.$$

Then for every $\rho \in \mathbb{R}$,

$$x = \sum_{i=1}^p \lambda_i x_i - \rho \sum_{i=1}^p \mu_i x_i = \sum_{i=1}^p (\lambda_i - \rho \mu_i) x_i. \quad (6.6)$$

Define the set

$$S = \{\sigma \in \mathbb{R} \mid \sigma \mu_i \leq \lambda_i \text{ for } 1 \leq i \leq p\}.$$

Then S is a closed interval of the form $[\alpha, \beta]$, as the following argument shows. For each i such that $\mu_i \neq 0$, σ must be in the interval $[\lambda_i/\mu_i, \infty)$ if $\mu_i < 0$ or $(-\infty, \lambda_i/\mu_i]$ if $\mu_i > 0$. Denote the appropriate interval by I_i . Then σ belongs to the intersection of all the intervals I_i which correspond to $\mu_i \neq 0$. Since the μ_i are not all zero and $\sum_{i=1}^p \mu_i = 0$, there exist integers m and n in $\{1, \dots, p\}$ such that $\mu_m < 0$ and $\mu_n > 0$. So the intersection is a closed bounded interval $[\alpha, \beta]$, where $\alpha = \lambda_j/\mu_j$ and $\beta = \lambda_k/\mu_k$ for some j and k in $\{1, \dots, p\}$.

Substituting α for ρ in the above representation (6.6) for x gives

$$x = \sum_{i=1}^p (\lambda_i - \alpha \mu_i) x_i,$$

where $(\lambda_i - \alpha \mu_i) \geq 0$ for $i = 1, \dots, p$, and $\sum_{i=1}^p (\lambda_i - \alpha \mu_i) = 1$. Since $\alpha = \lambda_j/\mu_j$, then $\lambda_j - \alpha \mu_j = 0$. Thus x is a convex combination of $p - 1$ points of A . Rename these $p - 1$ points x_1, \dots, x_{p-1} . If there exist $\mu_1, \dots, \mu_{p-1} \in \mathbb{R}$, not all zero, such that $\sum_{i=1}^{p-1} \mu_i x_i = 0$ and $\sum_{i=1}^{p-1} \mu_i = 0$, then the above process can be repeated to produce a representation of x as a convex combination of $p - 2$ points of A . Continuing this will eventually produce a representation of x as a convex combination of q points of A , where $q \leq n + 1$. To see this, suppose there exists a representation of x as a convex combination of q points x_1, \dots, x_q for $q > n + 1$. Suppose also that

whenever μ_1, \dots, μ_q are real numbers such that $\sum_{i=1}^q \mu_i x_i = 0$ and $\sum_{i=1}^q \mu_i = 0$, then $\mu_1 = \mu_2 = \dots = \mu_q = 0$. Then for any j , the set of $q - 1$ vectors $x_i - x_j$, for $i \neq j$ is linearly independent. But $q - 1 > n$, contradicting the fact that any linearly independent set in \mathbb{R}^n contains at most n vectors. \square

Before defining the moment space, it will be useful to state the following definition and theorems due to E. Helly. Proofs of the theorems are given in [24].

Definition 6.7 A set \mathcal{F} of functions is said to be uniformly of bounded variation on $[a, b]$ if there exists a constant M such that $V(f) = \int_a^b |df| \leq M$ for all $f \in \mathcal{F}$.

Theorem 6.9 (Helly's Selection Principle) Let $\{\theta_{m,n}\}$ be a double sequence of real numbers which is bounded by A ; i.e.,

$$|\theta_{m,n}| < A \quad \text{for all } m, n.$$

Then there exists a subsequence $\{n_k\}$ and a sequence $\{\theta_m\}$ of real numbers such that for every positive integer m ,

$$\lim_{k \rightarrow \infty} \theta_{m, n_k} = \theta_m$$

Theorem 6.10 (Helly) If $\{\sigma_n\}$ is a sequence of functions, uniformly of bounded variation on $[a, b]$ such that $\sigma_n(a)$ is bounded for all n , then there exists a subsequence $\{\sigma_{n_k}\}$ and a function σ of bounded variation such that $\lim_{k \rightarrow \infty} \sigma_{n_k}(x) = \sigma(x)$ for all x in $[a, b]$.

Definition 6.8 Let $\{u_i\}_0^n$ be a T -system on the interval $[a, b]$. The moment space \mathcal{M}_{n+1} with respect to $\{u_i\}_0^n$ is defined to be the set

$$\begin{aligned} \mathcal{M}_{n+1} = \{ & c = (c_0, c_1, \dots, c_n) \in E^{n+1} \mid c_i = \int_a^b u_i(t) d\sigma(t), i = 0, 1, \dots, n, \\ & \text{and } \sigma \text{ is a nondecreasing function of bounded variation} \\ & \text{which is right continuous on } (a, b) \}. \end{aligned}$$

Theorem 6.11 The moment space \mathcal{M}_{n+1} is a closed convex cone.

Proof: If λ is a positive real number and c is an element of \mathcal{M}_{n+1} , then λc also belongs to \mathcal{M}_{n+1} . So \mathcal{M}_{n+1} is a cone. Also, for any c_1 and c_2 in \mathcal{M}_{n+1} , $tc_1 + (1-t)c_2$ belongs to \mathcal{M}_{n+1} for all t satisfying $0 \leq t \leq 1$. So \mathcal{M}_{n+1} is convex.

To show that \mathcal{M}_{n+1} is closed, let $u(t) = \sum_{i=0}^n a_i u_i(t)$ denote a strictly positive u -polynomial. The existence of such a polynomial can be demonstrated using Theorem 6.4. (To construct one, find two nonnegative u -polynomials that have no common zeros. Then their sum is a strictly positive u -polynomial.) Let $c = (c_0, \dots, c_n) \in E^{n+1}$ and $\{c^{(r)}\}$ a sequence in \mathcal{M}_{n+1} with $\lim_{r \rightarrow \infty} c^{(r)} = c$. For each r , $c^{(r)} = (c_0^{(r)}, c_1^{(r)}, \dots, c_n^{(r)})$, where

$$c_i^{(r)} = \int_a^b u_i(t) d\sigma_r(t),$$

and σ_r is a nondecreasing function of bounded variation which is right continuous on (a, b) . For convenience, σ_r can be chosen so that $\sigma_r(a) = 0$ since $\int_a^b f d\sigma = \int_a^b f d(\sigma + K)$ for every function σ of bounded variation and every constant K . Since $\{c^{(r)}\}$ is convergent, the sequence $\{\sum_{i=0}^n a_i c_i^{(r)}\}$ is bounded. So for some constant M ,

$$M \geq \sum_{i=0}^n a_i c_i^{(r)} = \int_a^b u(t) d\sigma_r(t) \geq \left(\min_{a \leq \tau \leq b} u(\tau) \right) \int_a^b d\sigma_r(t),$$

where $\min_{a \leq \tau \leq b} u(\tau) > 0$. Since σ_r is nondecreasing, the variation of σ_r on $[a, b]$ is $\int_a^b d\sigma_r(t)$. Therefore, the sequence $\{\sigma_r\}$ is uniformly of bounded variation on $[a, b]$. By Helly's theorem, there exists a subsequence $\{\sigma_{r_k}\}$ and a function σ of bounded variation such that $\lim_{k \rightarrow \infty} \sigma_{r_k}(t) = \sigma(t)$ for every t in $[a, b]$. Then for each $i = 0, 1, \dots, n$, the continuity of u_i on $[a, b]$ implies

$$c_i = \lim_{k \rightarrow \infty} \int_a^b u_i(t) d\sigma_{r_k}(t) = \int_a^b u_i(t) d\sigma(t).$$

For a proof of this equality, see [24] or [38]. Since each σ_{r_k} is nondecreasing, and $\sigma_{r_k} \rightarrow \sigma$, then σ is also nondecreasing. Since σ is of bounded variation on $[a, b]$, it has at most a countable number of discontinuities, $\lim_{t \rightarrow x^-} \sigma(t)$ exists for each $x \in (a, b]$, and $\lim_{t \rightarrow x^+} \sigma(t)$ exists for each $x \in [a, b)$. (For verification of this, see [38].) Define the function σ_0 by $\sigma_0(a) = \sigma(a)$, $\sigma_0(b) = \sigma(b)$, and $\sigma_0(x) = \lim_{t \rightarrow x^+} \sigma(t)$ for all t in (a, b) . Then $\sigma = \sigma_0$ almost everywhere in $[a, b]$, and σ_0 is right continuous on (a, b) and nondecreasing on $[a, b]$. So σ_0 belongs to \mathcal{M}_{n+1} , and

$$c_i = \int_a^b u_i d\sigma = \int_a^b u_i d\sigma_0$$

for each $i = 0, 1, \dots, n$. Therefore $c \in \mathcal{M}_{n+1}$. □

Another characterization of \mathcal{M}_{n+1} results from the following definition.

Definition 6.9 Let $\{u_i\}_0^n$ be a T -system on $[a, b]$ and C_{n+1} the curve in E^{n+1} defined by

$$C_{n+1} = \{\gamma(t) = (u_0(t), u_1(t), \dots, u_n(t)) \mid a \leq t \leq b\}.$$

Let $\mathcal{C}(C_{n+1})$ denote the smallest convex cone containing C_{n+1} .

Theorem 6.12 $\mathcal{C}(C_{n+1})$ is closed and every $\gamma = (\gamma_0, \dots, \gamma_n) \in \mathcal{C}(C_{n+1})$ can be represented in the form

$$\gamma_i = \sum_{j=1}^{n+2} \lambda_j u_i(t_j), \quad i = 0, 1, \dots, n, \quad (6.7)$$

where for each $j = 1, \dots, n+2$, $\lambda_j \geq 0$, and $a \leq t_j \leq b$.

Proof: Clearly, all vectors of the form (6.7) belong to $\mathcal{C}(C_{n+1})$. Since $\mathcal{C}(C_{n+1})$ is convex, it must contain the convex hull of C_{n+1} . In fact, $\mathcal{C}(C_{n+1}) = \{k\beta \mid k \geq 0 \text{ and } \beta \text{ belongs to the convex hull of } C_{n+1}\}$. Let $\gamma \in \mathcal{C}(C_{n+1})$. Then $\gamma = k\beta$ for some $k \geq 0$ and β in the convex hull of C_{n+1} . By the theorem of Carathéodory, β can be represented in the form $\beta_i = \sum_{j=1}^{n+2} \alpha_j u_i(t_j)$, where $\alpha_j \geq 0$ and $a \leq t_j \leq b$. Then γ has the representation $\gamma_i = \sum_{j=1}^{n+2} k\alpha_j u_i(t_j)$, where $k\alpha_j \geq 0$.

To show that $\mathcal{C}(C_{n+1})$ is closed, suppose $\lim_{r \rightarrow \infty} \gamma^{(r)} = \gamma$, where $\gamma^{(r)} \in \mathcal{C}(C_{n+1})$ for every r . If $\gamma^{(r)} = (\gamma_0^{(r)}, \dots, \gamma_n^{(r)})$ and $\gamma = (\gamma_0, \dots, \gamma_n)$, then for each $i = 0, 1, \dots, n$,

$$\gamma_i^{(r)} = \sum_{j=1}^{n+2} \lambda_j^{(r)} u_i(t_j^{(r)}) \text{ and } \lim_{r \rightarrow \infty} \gamma_i^{(r)} = \gamma_i,$$

where for every r , $\lambda_j^{(r)} \geq 0$ and $a \leq t_j^{(r)} \leq b$. As seen before, there exists a strictly positive u -polynomial $u(t) = \sum_{i=0}^n a_i u_i(t)$. Since $\{\gamma^{(r)}\}$ is convergent, there exists M , such that for every positive integer r ,

$$\begin{aligned} M &\geq \sum_{i=0}^n a_i \gamma_i^{(r)} \\ &= \sum_{i=0}^n a_i \left(\sum_{j=1}^{n+2} \lambda_j^{(r)} u_i(t_j^{(r)}) \right) \\ &= \sum_{j=1}^{n+2} \left(\lambda_j^{(r)} \sum_{i=0}^n a_i u_i(t_j^{(r)}) \right) \\ &\geq \sum_{j=1}^{n+2} \lambda_j^{(r)} \left(\min_{a \leq \tau \leq b} u(\tau) \right). \end{aligned}$$

So $\{\lambda_j^{(r)}\}$ is uniformly bounded by M . Since each u_i is continuous on $[a, b]$, $\{|u_i(t_j^{(r)})|\}$ is uniformly bounded.

For each positive integer r , consider the finite sequence

$$\begin{aligned} \{\theta_{m,r}\} = & \{\lambda_1^{(r)}, \lambda_2^{(r)}, \dots, \lambda_{n+2}^{(r)}, \\ & u_0(t_1^{(r)}), u_0(t_2^{(r)}), \dots, u_0(t_{n+2}^{(r)}), \\ & u_1(t_1^{(r)}), u_1(t_2^{(r)}), \dots, u_1(t_{n+2}^{(r)}), \\ & \vdots \\ & u_n(t_1^{(r)}), u_n(t_2^{(r)}), \dots, u_n(t_{n+2}^{(r)})\}. \end{aligned}$$

Then $\{|\theta_{m,r}|\}$ is uniformly bounded, and Helly's selection principle guarantees the existence of a subsequence $\{r_k\}$ and sequences $\{\lambda_j\}$ and $\{u_{ij}\}$ such that

$$\begin{aligned} \lim_{k \rightarrow \infty} \lambda_j^{(r_k)} &= \lambda_j \quad j = 1, \dots, n+2 \quad \text{and} \\ \lim_{k \rightarrow \infty} u_i(t_j^{(r_k)}) &= u_{ij} \quad i = 0, \dots, n \text{ and } j = 1, \dots, n+2. \end{aligned}$$

Since $\lambda_j^{(r_k)} \geq 0$ for all j and r_k , then $\lambda_j \geq 0$. Since each u_i is continuous on a closed interval, the range is closed, and therefore u_{ij} is in the range of u_i . So for each i and j , there exists at least one $t_j^* \in [a, b]$ such that

$$\lim_{k \rightarrow \infty} u_i(t_j^{(r_k)}) = u_i(t_j^*).$$

This gives $\gamma = \lim_{r \rightarrow \infty} \gamma^{(r)} = (\gamma_0, \dots, \gamma_n)$ where for each i ,

$$\gamma_i = \lim_{k \rightarrow \infty} \gamma_i^{(r_k)} = \sum_{j=1}^{n+2} \lambda_j u_i(t_j^*).$$

So $\gamma \in \mathcal{C}(C_{n+1})$, and $\mathcal{C}(C_{n+1})$ is closed. □

In the proofs of some of the following theorems, some properties of hyperplanes will be useful. For convenience, the related definitions and theorems will be stated here. Proofs can be found in [40].

Definition 6.10 A hyperplane H in E^n is determined by a nonzero linear functional $f : E^n \rightarrow \mathbb{R}$, where $f(x_1, \dots, x_n) = \sum_{i=1}^n a_i x_i$ for real constants a_1, \dots, a_n , and a real constant d . H is defined by

$$H = \{v \in E^n | f(v) = d\} = f^{-1}(d).$$

Let $A, B \subset E^n$. Then H is said to separate A and B if either $f(A) \leq d$ and $f(B) \geq d$, or $f(A) \geq d$ and $f(B) \leq d$. H is said to separate A and B strictly if either $f(A) < d$ and $f(B) > d$ or $f(A) > d$ and $f(B) < d$.

Definition 6.11 Let $V \subset E^n$ and c an element of the boundary of V . A hyperplane $H = f^{-1}(d)$ in E^n is said to be a supporting hyperplane for V at c if $c \in H$ and either $f(V) \geq d$ or $f(V) \leq d$. A supporting hyperplane H for V is said to be non-trivial if V is not contained in H .

Theorem 6.13 Let V be a convex subset of E^n and c an element of the boundary of V . Then there exists a non-trivial supporting hyperplane for V at c .

Theorem 6.14 Let $V \subset E^n$ be closed, convex, and non-empty. Let $c \in E^n$ and $c \notin V$. Then there exists a hyperplane in E^n separating V and c strictly.

With these theorems, it can now be shown that \mathcal{M}_{n+1} and $\mathcal{C}(C_{n+1})$ are identical.

Theorem 6.15 $\mathcal{M}_{n+1} = \mathcal{C}(C_{n+1}) =$ the convex conical hull of C_{n+1} .

Proof: If $\gamma \in \mathcal{C}(C_{n+1})$, then $\gamma = (\gamma_0, \dots, \gamma_n)$ can be represented in the form (6.7). So for a given integer i , γ_i is given by

$$\gamma_i = \sum_{j=1}^{n+2} \lambda_j u_i(t_j)$$

where $t_1 \leq t_2 \leq \dots \leq t_{n+2}$, $a \leq t_j \leq b$, and $\lambda_j \geq 0$. There exists a nondecreasing step function σ which is right continuous on (a, b) and has its only points of increase at the points t_j and for which

$$\sum_{j=1}^{n+2} \lambda_j u_i(t_j) = \int_a^b u_i(t) d\sigma(t).$$

So $\gamma \in \mathcal{M}_{n+1}$.

Next, suppose $c^0 = (c_1^0, \dots, c_n^0) \in \mathcal{M}_{n+1}$ but $c^0 \notin \mathcal{C}(C_{n+1})$. Since $\mathcal{C}(C_{n+1})$ is a closed convex cone, Theorem 6.14 guarantees the existence of a hyperplane strictly separating c^0 from $\mathcal{C}(C_{n+1})$. So there exist real constants a_0, \dots, a_n , not all zero, and a real constant d such that

$$\sum_{i=0}^n a_i c_i^0 + d < 0,$$

and for every $\gamma = (\gamma_0, \dots, \gamma_n) \in \mathcal{C}(C_{n+1})$,

$$\sum_{i=0}^n a_i \gamma_i + d > 0.$$

For each nonnegative λ , $(\lambda u_0(t), \dots, \lambda u_n(t)) \in \mathcal{C}(C_{n+1})$ for all $t \in [a, b]$. Therefore

$$\sum_{i=0}^n a_i \lambda u_i(t) + d > 0. \quad (6.8)$$

Since $c^0 \in \mathcal{M}_{n+1}$, there exists a nondecreasing function σ^0 which is right continuous on (a, b) and satisfies

$$\begin{aligned} 0 &> \sum_{i=0}^n a_i c_i^0 + d \\ &= \sum_{i=0}^n a_i \int_a^b u_i(t) d\sigma^0(t) + d \\ &= \int_a^b \left(\sum_{i=0}^n a_i u_i(t) \right) d\sigma^0(t) + d. \end{aligned} \quad (6.9)$$

Denote $\int_a^b d\sigma^0(t)$ by λ . Since $c^0 \notin \mathcal{C}(C_{n+1})$, $c^0 \neq (0, \dots, 0)$. So $\lambda > 0$, and integrating (6.8) with respect to σ^0 gives

$$\begin{aligned} 0 &\leq \int_a^b \left(\sum_{i=0}^n a_i \lambda u_i(t) + d \right) d\sigma^0(t) \\ &= \int_a^b \left(\sum_{i=0}^n a_i \lambda u_i(t) \right) d\sigma^0(t) + \int_a^b d d\sigma^0(t) \\ &= \lambda \int_a^b \left(\sum_{i=0}^n a_i u_i(t) \right) d\sigma^0(t) + \lambda d. \end{aligned}$$

Dividing by λ gives

$$0 \leq \int_a^b \left(\sum_{i=0}^n a_i u_i(t) \right) d\sigma^0(t) + d.$$

This contradicts (6.9), so $c^0 \in \mathcal{C}(C_{n+1})$, and $\mathcal{M}_{n+1} = \mathcal{C}(C_{n+1})$. \square

Definition 6.12 The index $I(c)$ of a point c in \mathcal{M}_{n+1} is defined to be the minimal number of points of C_{n+1} that can be used in a convex representation of c under the special convention that $(u_0(a), u_1(a), \dots, u_n(a))$ and $(u_0(b), u_1(b), \dots, u_n(b))$ are counted as half points while $(u_0(t), u_1(t), \dots, u_n(t))$ receives a full count for $a < t < b$.

Theorem 6.16 *A vector $c^0 \in \mathcal{M}_{n+1}$, $c^0 \neq 0$, is a boundary point of \mathcal{M}_{n+1} if and only if $I(c^0) < (n+1)/2$. Moreover, every boundary point c^0 admits a unique representation*

$$c_i^0 = \sum_{j=1}^p \lambda_j u_i(t_j), \quad i = 0, 1, \dots, n \quad (6.10)$$

where $p \leq \frac{n+2}{2}$, $\lambda_j > 0$, $j = 1, 2, \dots, p$, and $t_1 < t_2 < \dots < t_p$.

Proof: First, we let $c^0 \neq 0$ be a boundary point of \mathcal{M}_{n+1} and show that $I(c_0) < \frac{n+1}{2}$. Since \mathcal{M}_{n+1} is closed, $c^0 \in \mathcal{M}_{n+1}$, and by Theorem 6.13, there exists a supporting hyperplane to \mathcal{M}_{n+1} at c^0 . That is, there exist real constants a_0, \dots, a_n , not all zero, and a real constant d for which

$$\sum_{i=0}^n a_i c_i + d \geq 0 \quad \text{for all } c = (c_0, \dots, c_n) \in \mathcal{M}_{n+1} \quad (6.11)$$

$$\text{and} \quad \sum_{i=0}^n a_i c_i^0 + d = 0 \quad \text{for } c^0 = (c_0^0, \dots, c_n^0) \quad (6.12)$$

Since $(0, \dots, 0) \in \mathcal{M}_{n+1}$, (6.11) gives $d \geq 0$. Suppose $d > 0$. Then (6.12) implies $\sum_{i=0}^n a_i c_i^0 < 0$. So there exists a positive real number λ such that

$$\lambda \left(\sum_{i=0}^n a_i c_i^0 \right) + d < 0. \quad (6.13)$$

But $\lambda c^0 \in \mathcal{M}_{n+1}$ for each positive real number λ , so by (6.11),

$$\begin{aligned} \sum_{i=0}^n a_i (\lambda c_i^0) + d &\geq 0 \\ \text{or } \lambda (\sum_{i=0}^n a_i c_i^0) + d &\geq 0 \end{aligned}$$

This contradicts (6.13), so $d = 0$, and we have

$$\sum_{i=0}^n a_i c_i \geq 0 \quad \text{for all } c \in \mathcal{M}_{n+1}, \text{ and} \quad (6.14)$$

$$\sum_{i=0}^n a_i c_i^0 = 0.$$

Define the function u^0 by $u^0(t) = \sum_{i=0}^n a_i u_i(t)$. For each $t \in [a, b]$, $(u_0(t), \dots, u_n(t)) \in \mathcal{M}_{n+1}$. To see this, choose the step function σ_i which is right-continuous and has its only point of increase at t of magnitude 1. Then for each i ,

$$u_i(t) = \int_a^b u_i(\tau) d\sigma_i(\tau).$$

So $u^0(t) \geq 0$ by (6.14). Since $c^0 \in \mathcal{M}_{n+1}$, there exists a function σ^0 such that $c_i^0 = \int_a^b u_i(t) d\sigma^0(t)$ for each i . Then $\int_a^b u^0(t) d\sigma^0(t) = 0$. Since $u^0(t) \geq 0$ for all $t \in [a, b]$ and σ^0 is nondecreasing, then u^0 vanishes at every point of increase of σ^0 . So σ^0 can only have a point of increase at a zero of u^0 . By Theorem 6.2, the number of zeros of u^0 is less than or equal to n , and they will be denoted by t_1, t_2, \dots, t_k , where $k \leq n$. Then for some positive constants $\lambda_1, \dots, \lambda_k$,

$$\int_a^b u_i(t) d\sigma^0(t) = \sum_{j=1}^k \lambda_j u_i(t_j)$$

for each i . So c^0 has the representation

$$c^0 = \sum_{j=1}^k \lambda_j (u_0(t_j), \dots, u_n(t_j)).$$

Since $u^0(t) \geq 0$ on $[a, b]$, all zeros except a and b are nonnodal. So $2I(c) \leq \tilde{Z}(u^0)$. (See Definition 6.5.) By Theorem 6.3, $\tilde{Z}(u^0) \leq n$, so $I(c) < \frac{n+1}{2}$.

To show that the representation (6.10) is unique, let $T = \{\tau_j\}_0^n$ be a set of distinct points in $[a, b]$ containing all the zeros of $u^0(t)$. If c^0 has a representation of the form (6.10), then there exists a function σ^1 such that $c^0 = (\int_a^b u_0(t) d\sigma^1(t), \dots, \int_a^b u_n(t) d\sigma^1(t))$, and the points of increase of σ^1 are zeros of $u^0(t)$. The system

$$\begin{aligned} c_0^0 &= \beta_0 u_0(\tau_0) + \dots + \beta_n u_0(\tau_n) \\ &\vdots \\ c_n^0 &= \beta_0 u_n(\tau_0) + \dots + \beta_n u_n(\tau_n) \end{aligned}$$

has a unique solution for β_0, \dots, β_n since the determinant of coefficients is nonzero.

Now let c^0 denote a vector of \mathcal{M}_{n+1} for which $I(c^0) < \frac{n+1}{2}$. Then $c^0 = (c_0^0, \dots, c_n^0)$ has the representation $c_i^0 = \sum_{j=1}^p \lambda_j u_i(t_j)$, where $\sum_{j=1}^p \omega(t_j) = 2I(c^0) < n+1$. (See Definition 6.6.) Then $\sum_{j=1}^p \omega(t_j) \leq n$ and, by Theorem 6.4, there exists a nontrivial, nonnegative u -polynomial $u^0(t) = \sum_{i=0}^n a_i u_i(t)$ for which $u^0(t_j) = 0$ for each t_j . If $c = (c_0, \dots, c_n) \in \mathcal{C}(C_{n+1})$, then by Theorem 6.12, c has the representation

$$c = \left(\sum_{j=1}^{n+2} \delta_j u_0(\tau_j), \dots, \sum_{j=1}^{n+2} \delta_j u_n(\tau_j) \right)$$

where each $\delta_j \geq 0$. Then

$$\begin{aligned} a_0 c_0 + \cdots + a_n c_n &= a_0 \sum_{j=1}^{n+2} \delta_j u_0(\tau_j) + \cdots + a_n \sum_{j=1}^{n+2} \delta_j u_n(\tau_j) \\ &= \delta_1 (a_0 u_0(\tau_1) + \cdots + a_n u_n(\tau_1)) + \cdots + \\ &\quad \delta_{n+2} (a_0 u_0(\tau_{n+2}) + \cdots + a_n u_n(\tau_{n+2})) \geq 0. \end{aligned}$$

Also,

$$\begin{aligned} a_0 c_0^0 + \cdots + a_n c_n^0 &= a_0 \sum_{j=1}^p \lambda_j u_0(t_j) + \cdots + a_n \sum_{j=1}^p \lambda_j u_n(t_j) \\ &= \lambda_1 [a_0 u_0(t_1) + \cdots + a_n u_n(t_1)] + \cdots + \\ &\quad \lambda_p [a_0 u_0(t_p) + \cdots + a_n u_n(t_p)] = 0. \end{aligned}$$

So the points satisfying $a_0 c_0 + \cdots + a_n c_n = 0$ form the supporting hyperplane to \mathcal{M}_{n+1} at c^0 . Therefore, c^0 is on the boundary of \mathcal{M}_{n+1} . \square

Before looking at interior points of \mathcal{M}_{n+1} , it is helpful to define a section of \mathcal{M}_{n+1} .

Definition 6.13 A section of \mathcal{M}_{n+1} is any subset S of \mathcal{M}_{n+1} with the following properties:

1. S is contained in a hyperplane.
2. If $c^0 \in \mathcal{M}_{n+1}$ and $c^0 \neq 0$, then there exists a unique positive real number λ such that $\lambda c^0 \in S$.

Theorem 6.17 S is a section in \mathcal{M}_{n+1} if and only if there exist real constants a_0, \dots, a_n and a positive constant α such that $u(t) = \sum_{i=0}^n a_i u_i(t) > 0$ for all $t \in [a, b]$, and

$$S = \{(c_0, \dots, c_n) \in \mathcal{M}_{n+1} \mid \sum_{i=0}^n a_i c_i = \alpha\} \quad (6.15)$$

Proof: First, suppose $S = \{(c_0, \dots, c_n) \in \mathcal{M}_{n+1} \mid \sum_{i=0}^n a_i c_i = \alpha\}$, where $\sum_{i=0}^n a_i u_i(t) > 0$ for all $t \in [a, b]$, and $\alpha > 0$. Let $c \in S$. Then clearly c belongs to the hyperplane defined by $\sum_{i=0}^n a_i x_i = \alpha$.

Let $c = (c_0, \dots, c_n) \in \mathcal{M}_{n+1}$ and $c \neq 0$. By theorem (6.8) c has the representation

$$c = \left(\sum_{j=1}^{n+2} \lambda_j u_0(t_j), \dots, \sum_{j=1}^{n+2} \lambda_j u_n(t_j) \right)$$

where $\lambda_j \geq 0$ and $a \leq t_j \leq b$ for $j = 1, \dots, n+2$. So

$$\begin{aligned} \sum_{i=0}^n a_i c_i &= \sum_{i=0}^n \left(a_i \sum_{j=1}^{n+2} \lambda_j u_i(t_j) \right) \\ &= \sum_{j=1}^{n+2} \left(\lambda_j \sum_{i=0}^n a_i u_i(t_j) \right). \end{aligned}$$

Since $u(t) > 0$ for all $t \in [a, b]$, $\sum_{i=0}^n a_i u_i(t_j) > 0$. Since $c \neq 0$, we have $\lambda_j \neq 0$ for some j and therefore $\sum_{i=0}^n a_i c_i > 0$. So there exists a unique positive constant λ such that $\sum_{i=0}^n a_i (\lambda c_i) = \alpha$, and $\lambda c \in S$. This shows that S is a section.

Now suppose S is a section in \mathcal{M}_{n+1} . Then S lies in a hyperplane defined by $\sum_{i=0}^n a_i x_i = \alpha$, where $\alpha \geq 0$ and $\sum_{i=0}^n a_i^2 > 0$. The following argument shows that $\alpha > 0$. Suppose $\alpha = 0$, and let $c \in \mathcal{M}_{n+1}$. Then there exists a positive real number λ such that $\lambda c \in S$. If $c = (c_0, \dots, c_n)$, then $\sum_{i=0}^n a_i \lambda c_i = 0$ and therefore $\sum_{i=0}^n a_i c_i = 0$. For every $t \in [a, b]$, $(u_0(t), \dots, u_n(t)) \in \mathcal{M}_{n+1}$ and thus $\sum_{i=0}^n a_i u_i(t) \equiv 0$ on $[a, b]$. By Theorem 6.2, the number of distinct zeros of a non-trivial u -polynomial is less than or equal to n . So $\sum_{i=0}^n a_i^2 = 0$, which contradicts the hypothesis that $\sum_{i=0}^n a_i^2 > 0$. So $\alpha > 0$. For any $t \in [a, b]$, there exists $\lambda > 0$ such that $\lambda(u_0(t), \dots, u_n(t)) \in S$. Since S lies in the hyperplane defined by $\sum_{i=0}^n a_i x_i = \alpha$, we have $\sum_{i=0}^n a_i \lambda u_i(t) = \alpha$ and thus $\sum_{i=0}^n a_i u_i(t) > 0$. Let $c = (c_0, \dots, c_n)$ be any element of \mathcal{M}_{n+1} for which $\sum_{i=0}^n a_i c_i = \alpha$. Since S is a section, there exists $\lambda \in \mathbb{R}$ such that $\lambda c \in S$. Then $\sum_{i=0}^n a_i \lambda c_i = \alpha$ and therefore $\lambda = 1$. So $c \in S$, and

$$\{(c_0, \dots, c_n) \in \mathcal{M}_{n+1} \mid \sum_{i=0}^n a_i c_i = \alpha\} \subset S.$$

Therefore $S = \{(c_0, \dots, c_n) \in \mathcal{M}_{n+1} \mid \sum_{i=0}^n a_i c_i = \alpha\}$. □

Theorem 6.18 *Let S be a section in \mathcal{M}_{n+1} . Then S is convex and bounded.*

Proof: If S is a section in \mathcal{M}_{n+1} , then by the previous theorem, there exist real constants a_0, \dots, a_n such that $u(t) = \sum_{i=0}^n a_i u_i(t) > 0$ for all $t \in [a, b]$, and $S = \{(c_0, \dots, c_n) \in \mathcal{M}_{n+1} \mid \sum_{i=0}^n a_i c_i = \alpha\}$ for $\alpha > 0$. Since $u(t) > 0$ on $[a, b]$, $\min_{a \leq t \leq b} u(t) = M$ for some positive number M . Let $c = (c_0, \dots, c_n) \in S$. Then $c \in \mathcal{M}_{n+1}$, and there exists a right continuous nondecreasing function σ such that

$c_i = \int_a^b u_i(t) d\sigma(t)$ for each i . So

$$\begin{aligned}\alpha &= \sum_{i=0}^n a_i c_i \\ &= \sum_{i=0}^n a_i \int_a^b u_i(t) d\sigma(t) \\ &= \int_a^b \left(\sum_{i=0}^n a_i u_i(t) \right) d\sigma(t) \\ &\geq M \int_a^b d\sigma(t).\end{aligned}$$

So $\int_a^b d\sigma(t) \leq \alpha/M$. Since u_i is continuous on $[a, b]$, it is bounded there, so

$$\begin{aligned}|c_i| &= \left| \int_a^b u_i(t) d\sigma(t) \right| \\ &\leq \int_a^b |u_i(t)| d\sigma(t) \\ &\leq \max_{a \leq t \leq b} |u_i(t)| \frac{\alpha}{M}.\end{aligned}$$

So S is bounded.

To show that S is convex, let u and v belong to S , and let $t \in [0, 1]$. By Theorem 6.11, $tu + (1-t)v \in \mathcal{M}_{n+1}$. By Theorem 6.17, there exist real constants a_0, \dots, a_n and a positive constant α such that

$$S = \{(c_0, \dots, c_n) \in \mathcal{M}_{n+1} \mid \sum_{i=0}^n a_i c_i = \alpha\}.$$

Let $u = (u_0, \dots, u_n)$ and $v = (v_0, \dots, v_n)$. Then

$$\sum_{i=0}^n a_i u_i = \alpha = \sum_{i=0}^n a_i v_i.$$

So

$$\begin{aligned}\alpha &= t \sum_{i=0}^n a_i u_i + (1-t) \sum_{i=0}^n a_i v_i \\ &= \sum_{i=0}^n a_i (tu_i + (1-t)v_i),\end{aligned}$$

and therefore $tu + (1-t)v$ belongs to S . □

Definition 6.14 Let $c = (c_0, \dots, c_n) \in \mathcal{M}_{n+1}$. The values $\{t_j\}$ in the representation

$$c_i = \sum_{j=1}^p \lambda_j u_i(t_j), \quad i = 0, 1, \dots, n; \quad \lambda_j > 0, \quad 1 \leq j \leq p$$

are called roots of the representation. If t^* is a root, we say that the representation involves t^* . The nondecreasing function σ which concentrates all its increase at the points of $\{t_j\}_1^p$, with respective weights $\{\lambda_j\}$ is referred to as the associated measure of the representation. (Note that $c_i = \int_a^b u_i(t) d\sigma(t) = \sum_{j=1}^p \lambda_j u_i(t_j)$.) The index of the set $T = \{t_1, \dots, t_p\}$ is defined to be that number obtained by counting interior points as one and the end points a and b as one half. The index of the representation of c and the index of the measure σ generating c are each defined to be the index of the set of roots of the representation.

Note that, if a representation of $c \in \mathcal{M}_{n+1}$ involves the minimum number of points, then the index of the representation is the same as the index of c , $I(c)$, in Definition 6.12.

Theorem 6.19 Let $c = (c_0, \dots, c_n)$ be an interior point of \mathcal{M}_{n+1} . For each $t^* \in [a, b]$, there exists a representation

$$c_i = \sum_{j=1}^p \lambda_j u_i(t_j) \quad \text{for } i = 0, 1, \dots, n; \quad \lambda_j > 0;$$

$$j = 1, 2, \dots, p; \quad t^* \in \{t_1, t_2, \dots, t_p\}$$

of index $(n+1)/2$ or $(n+2)/2$.

Proof: Let c be an interior point of \mathcal{M}_{n+1} and $t^* \in [a, b]$. Suppose there exists $\lambda > 0$ such that $c = \lambda(u_0(t^*), \dots, u_n(t^*))$. If $n = 1$ and either $t^* = a$ or $t^* = b$, then the index of c , $I(c)$, would be $\frac{1}{2}$. But by Theorem 6.16, c would be a boundary point. So $t^* \in (a, b)$, and c has a representation of index 1. In this case the conclusion of the theorem is satisfied. If $n \geq 2$, then the index of c , $I(c)$, satisfies $I(c) \leq 1 < \frac{n+1}{2}$. Again by Theorem 6.16, c would be a boundary point. So there does not exist $\lambda > 0$ such that $c = \lambda(u_0(t^*), \dots, u_n(t^*))$ in case $n \geq 2$.

Consider a section $S \in \mathcal{M}_{n+1}$ containing c . Let $c^* = \lambda(u_0(t^*), \dots, u_n(t^*))$, where $\lambda > 0$ and $c^* \in S$. In case $n = 1$ and $c^* = c$, the theorem holds, so assume $c \neq c^*$. In

case $n \geq 2$, it is necessary that $c^* \neq c$. Since S is bounded, the ray $c^* + t(c - c^*)$, $t \geq 0$, lies in S and pierces the boundary of \mathcal{M}_{n+1} at a point $\tilde{c} = c^* + \tau(c - c^*)$, where $\tau > 1$. So $c = \alpha\tilde{c} + (1 - \alpha)c^*$ for $\alpha = 1/\tau \in (0, 1)$.

If $I(\tilde{c}) < \frac{n-1}{2}$, then $I(c) < \frac{n-1}{2} + 1$, and by Theorem 6.16, c would be a boundary point of \mathcal{M}_{n+1} . So $I(\tilde{c}) \geq \frac{n-1}{2}$, and we have $\frac{n-1}{2} \leq I(\tilde{c}) < \frac{n+1}{2}$. Then $I(\tilde{c}) = \frac{n-1}{2}$ or $\frac{n}{2}$, and therefore $I(c) = \frac{n+1}{2}$ or $\frac{n+2}{2}$. Since α and $(1 - \alpha)$ are both positive, Theorem 6.16 guarantees that c has a representation

$$c_i = \sum_{j=1}^p \lambda_j u_i(t_j)$$

where $i = 0, \dots, n$, and $\lambda_j > 0$ for each $j = 1, \dots, p$. □

Definition 6.15 Let c be an interior point of \mathcal{M}_{n+1} . A representation for c of index $I(c) = (n+1)/2$ is called principal, and any representation of index $I(c) \leq (n+2)/2$ is called canonical. A canonical or principal representation is designated upper if it involves the end point b and lower if it does not.

Theorem 6.20 Let c be an interior point of \mathcal{M}_{n+1} for $n \geq 2$. There exist at least two principal representations. If $n = 2m$, one representation is found by prescribing $t^* = a$ in Theorem 6.19, and the other is found by prescribing $t^* = b$. The roots, respectively, are

$$a = t_1^* < t_2^* < \dots < t_{m+1}^* < b$$

$$a < s_1^* < s_2^* < \dots < s_{m+1}^* = b$$

If $n = 2m + 1$, one principal representation is found by prescribing $t^* = a$ to get roots

$$a = s_1^* < s_2^* < \dots < s_{m+2}^* = b$$

The other principal representation has roots

$$a < t_1^* < t_2^* < \dots < t_{m+1}^* < b$$

Proof: Let c be an interior point of \mathcal{M}_{n+1} . First suppose $n = 2m$ for some positive integer m . By Theorem 6.19, there exists a representation of index $\frac{n+1}{2}$ or $\frac{n+2}{2}$ involving $t^* = a$. In the proof of the theorem, it was shown that c has a representation of the form $c = \alpha\tilde{c} + (1 - \alpha)c^*$, where $0 < \alpha < 1$, $c^* = \lambda(u_0(a), \dots, u_n(a))$ for some

$\lambda > 0$, and \tilde{c} is on the boundary of \mathcal{M}_{n+1} . Since $I(\tilde{c}) < \frac{n+1}{2} = m + \frac{1}{2}$, there exists a representation of \tilde{c} of index less than or equal to m . If the representation of \tilde{c} is less than m , then the index of the representation $c = \alpha\tilde{c} + (1-\alpha)c^*$ is less than $m + \frac{1}{2}$. But this would mean c is a boundary point of \mathcal{M}_{n+1} . So $I(\tilde{c}) = m$. The representation of \tilde{c} does not involve a because if it did, the index of the representation of c would be equal to the index of the representation of the boundary point \tilde{c} . Since m is an integer, the representation of \tilde{c} cannot involve b . So the representation of c involving a has index $m + \frac{1}{2}$ and thus is principal. The representation does not involve b . A second principal representation not involving a is determined by letting $t^* = b$.

Now let $n = 2m + 1$ for a positive integer m . As above, let $t^* = a$, $c^* = \lambda(u_0(a), \dots, u_n(a))$ with $\lambda > 0$, and $c = \alpha\tilde{c} + (1-\alpha)c^*$ for $0 < \alpha < 1$. Then $I(\tilde{c}) < \frac{n+1}{2}$. So $I(\tilde{c}) \leq m + \frac{1}{2}$. If $I(\tilde{c}) < m + \frac{1}{2}$, then the given representation for c would have index less than $m + 1$, contradicting the fact that c is an interior point. So $I(\tilde{c}) = m + \frac{1}{2}$, and the representation must involve either a or b . It cannot involve a because, in that case, the representation of the interior point c would have the same index as that of the boundary point \tilde{c} . Thus the representation of c involves both a and b and has index $m + 1 = \frac{n+1}{2}$. So it is principal.

To construct another principal representation define the set

$$C_{n+1}(d) = \{(u_1(t), \dots, u_n(t)) \mid d \leq t \leq b\}$$

where $d \in (a, b)$. Let $\mathcal{M}_{n+1}(d)$ be the convex cone spanned by the curve $C_{n+1}(d)$. First, we show there exists $d' > a$ such that c belongs to the boundary of $\mathcal{M}_{n+1}(d')$. Suppose no such d' exists. Define the sets A_1 and A_2 by

$$\begin{aligned} A_1 &= \{d \mid c \text{ is an interior point of } \mathcal{M}_{n+1}(d)\} \\ A_2 &= \{d \mid c \notin \mathcal{M}_{n+1}(d)\} \end{aligned}$$

The set A_1 is not empty. To see this, note that since c is an interior point of \mathcal{M}_{n+1} , Theorem 6.20 guarantees it has a principal representation not involving a . Let t^* be the smallest root, and choose d to satisfy $a < d < t^*$. Then c is an interior point of $\mathcal{M}_{n+1}(d)$ because the index of c with respect to $[a, b]$ is the same as the index with respect to $[d, b]$.

The set A_2 can be shown to be nonempty by considering separately the cases where n is even and where n is odd. First, choose a principal representation of c with

respect to $[a, b]$ which does not involve b . Denoting $(u_0(t), \dots, u_n(t))$ by $U(t)$, c can be expressed as

$$c = \sum_{j=1}^p \lambda_j U(t_j),$$

where each λ_j is positive. Let t_p be the largest root, and choose $d = t_p$.

First, suppose $n = 2m$ for a positive integer m . Then the above principal representation involves the endpoint a , and $p = m + 1$. If $c \in \mathcal{M}_{n+1}(d)$, then it has a representation involving d of index less than or equal to $\frac{n+1}{2}$. Let

$$c = \sum_{j=1}^q \gamma_j U(s_j),$$

where $s_1 = d$, $q \leq m + 1$, and each γ_j is positive. Then

$$\sum_{j=1}^p \lambda_j U(t_j) - \sum_{j=1}^q \gamma_j U(s_j) = 0.$$

Since $t_p = s_1$, we have

$$\sum_{j=1}^k \alpha_j U(r_j) = 0,$$

where $\{r_1, \dots, r_k\}$ is a set of k distinct points in $[a, b]$, and $k \leq 2m + 1 = n + 1$. Since $\{u_0(t), \dots, u_n(t)\}$ is a T-System, this would mean each $\alpha_j = 0$ for $j = 1, \dots, k$. But this contradicts the assumption that each λ_j in the principal representation of c with respect to $[a, b]$ is positive.

Now suppose $n = 2m + 1$ for m a positive integer. Since the representation of c does not involve b , then by Theorem 6.20, it does not involve a . So $p = m + 1$. If $c \in \mathcal{M}_{n+1}(d)$, then c has a representation in $\mathcal{M}_{n+1}(d)$ of index less than or equal to $m + 1$. If the index is equal to $m + 1$, then c is an interior point of $\mathcal{M}_{n+1}(d)$, and the representation can be selected so that neither d nor b is involved. So c has a representation

$$c = \sum_{j=1}^q \gamma_j U(s_j),$$

where $q \leq m + 1$. Then

$$\begin{aligned} 0 &= \sum_{j=1}^p \lambda_j U(t_j) - \sum_{j=1}^q \gamma_j U(s_j) \\ &= \sum_{j=1}^k \alpha_j U(r_j), \end{aligned}$$

where $k \leq 2(m+1) = n+1$. As before a contradiction is obtained because $\{u_0(t), \dots, u_n(t)\}$ is a T-System. So both A_1 and A_2 are nonempty, and they can be shown to be open. Suppose c does not belong to the boundary of $\mathcal{M}_{n+1}(d)$ for any $d > a$. Then A_1 and A_2 provide a decomposition of the interval (a, b) into the disjoint union of open sets. But this contradicts the fact that (a, b) is connected. Therefore, there exists $d' \in (a, b)$ such that c belongs to the boundary of $\mathcal{M}_{n+1}(d')$. By Theorem 6.16, since c is a boundary point of $\mathcal{M}_{n+1}(d')$, the index of c related to $[d', b]$ is less than or equal to $m + \frac{1}{2}$. If the index were less than or equal to m , then the index with respect to $[a, b]$ would be less than or equal to $m + \frac{1}{2}$. But this would contradict the fact that c is an interior point of \mathcal{M}_{n+1} . So the index of the representation related to $[d', b]$ must be $m + \frac{1}{2}$. If b is involved, then since m is a positive integer, d' is not involved. In this case, the index relative to $[a, b]$ would also be $m + \frac{1}{2}$, contradicting the fact that c is an interior point of \mathcal{M}_{n+1} . So b is not involved and therefore d' must be involved. This representation has $m+1$ roots all belonging to (a, b) and by definition is principal. \square

Theorem 6.21 *Let $c = (c_0, \dots, c_n)$ be an interior point of \mathcal{M}_{n+1} and let σ^* and σ represent two different measures satisfying $\int \mu_i d\sigma^* = \int \mu_i d\sigma = c_i$ ($i = 0, 1, \dots, n$) where σ has index $\frac{n+1}{2}$ or $\frac{n+2}{2}$. Let $T = \{t_1, t_2, \dots, t_p\}$ be the roots of σ , where $t_1 < t_2 < \dots < t_p$. Then for every pair of roots t_j and t_{j+1} of σ lying in the open interval (a, b) , there exists a point of increase of σ^* in the open interval (t_j, t_{j+1}) . If σ has index $\frac{n+1}{2}$, this remains true if $t_j = a$, or $t_{j+1} = b$.*

Proof: First, note that since σ and σ^* are measures, then they are right continuous non-decreasing functions which concentrate all their increase at their roots. Let t_j and t_{j+1} be consecutive roots of σ . Note that the existence of two distinct roots implies that $n \geq 1$. If the index of σ is $\frac{n+2}{2}$, assume these roots belong to the interval (a, b) . In this case n must be greater than 1, for if $n = 1$, then the index of σ would be $1\frac{1}{2}$. This would mean there are only two roots, and one is an endpoint. So there are not two distinct roots in (a, b) . If the index of σ is $\frac{n+1}{2}$, one of the roots can be an endpoint.

Suppose σ^* does not increase in the interval (t_j, t_{j+1}) . Let $T = \{t_1, t_2, \dots, t_p\}$ be the set of roots of σ and define a function $\omega : T \rightarrow \{1, 2\}$ by

$$\omega(t_i) = 1 \text{ if } t_i \in \{a, t_j, t_{j+1}, b\}$$

$$\omega(t_i) = 2 \text{ if } t_i \notin \{a, t_j, t_{j+1}, b\}.$$

Under the above assumptions, it follows that

$$\sum_{i=1}^p \omega(t_i) \leq n.$$

By Theorem 6.5, there exists a u -polynomial u such that $u(t) \neq 0$ for $t \in (a, b) - T$ and such that t_i is a nodal zero if $\omega(t_i) = 1$ and a nonnodal zero if $\omega(t_i) = 2$. Since $\omega(t_j) = 1 = \omega(t_{j+1})$, t_j and t_{j+1} are nodal zeros of u . So u must change signs at t_j if $t_j \neq a$ and at t_{j+1} if $t_{j+1} \neq b$. Then (after multiplying by -1 if necessary) u vanishes on (a, b) precisely at the roots of σ and has the following properties:

$$u(t) \begin{cases} \geq 0 & \text{if } t \notin [t_j, t_{j+1}] \\ < 0 & \text{if } t \in (t_j, t_{j+1}) \end{cases}$$

Now $u(t) = \sum_{i=0}^n a_i u_i(t)$ for constants a_0, a_1, \dots, a_n and by assumption $\int u_i d\sigma^* = \int u_i d\sigma$ for each $i = 1, 2, \dots, n$. Since $u(t) = 0$ at the roots of σ , then $\int u(t) d\sigma = 0$. Also, σ^* has no points of increase in (t_j, t_{j+1}) , so

$$\begin{aligned} 0 &= \int u(t)(d\sigma^* - d\sigma) = \int u(t) d\sigma^* \\ &= \int_{[a, t_j]} u(t) d\sigma^* + \int_{[t_{j+1}, b]} u(t) d\sigma^* \end{aligned}$$

Since $u(t) \geq 0$ for $t \notin [t_j, t_{j+1}]$, if σ^* increases at a point which is not a root of σ , then $\int_{[a, t_j]} u(t) d\sigma^* + \int_{[t_{j+1}, b]} u(t) d\sigma^* > 0$. This contradiction implies that the roots of σ^* are also roots of σ . If σ has index $\frac{n+2}{2}$, then as indicated above, n must be greater than 1, and the number of roots of σ is less than or equal to $\frac{n+2}{2} + 1$. If σ has index $\frac{n+1}{2}$, the number of roots is less than or equal to $\frac{n+1}{2} + 1$. In either case, the number of roots of σ is less than or equal to $n + 1$. Since $\{u_0, u_1, \dots, u_n\}$ is a T-system, σ and σ^* cannot be distinct. Therefore, if σ and σ^* are different representations, then σ^* must have a point of increase in the interval (t_j, t_{j+1}) . \square

Theorem 6.22 *For each c in the interior of \mathcal{M}_{n+1} , there exist precisely two principal representations. The roots of these representations strictly interlace.*

The proof of this theorem is given in [28].

Example: Let u_0, u_1, u_2 , and u_3 be the functions

$$u_0(t) = 1 \quad u_1(t) = t \quad u_2(t) = t^2 \quad u_3(t) = t^3$$

These functions form a T-System on $[0, 1]$, and \mathcal{M}_4 is the smallest convex cone containing the set

$$C_4 = \{(u_0(t), u_1(t), u_2(t), u_3(t)) \mid t \in [0, 1]\}$$

Let c be the point

$$c = (c_0, c_1, c_2, c_3) = \left(\int_0^1 1 \, dt, \int_0^1 t \, dt, \int_0^1 t^2 \, dt, \int_0^1 t^3 \, dt \right) = (1, 1/2, 1/3, 1/4).$$

By definition, $c \in \mathcal{M}_4$. Also, c is not a boundary point of \mathcal{M}_4 , so by Theorem 6.22, there exist precisely two principal representations of c , and the roots of the two representations strictly interlace. By Theorem 6.20, the two sets of roots must be $\{s_1, s_2, s_3\}$ and $\{t_1, t_2\}$, where

$$s_1 = 0, \quad 0 < s_2 < 1, \quad s_3 = 1 \quad \text{and}$$

$$0 < t_1 < t_2 < 1.$$

In the first case, the representation of $c = (c_0, c_1, c_2, c_3)$ is given by

$$c_i = \alpha_1 u_i(s_1) + \alpha_2 u_i(s_2) + \alpha_3 u_i(s_3)$$

where

$$s_1 = 0, \quad s_2 = 1/2, \quad s_3 = 1$$

$$\alpha_1 = 1/6, \quad \alpha_2 = 2/3, \quad \alpha_3 = 1/6.$$

The second representation is found by Gaussian quadrature to be

$$c_i = \beta_1 u_i(t_1) + \beta_2 u_i(t_2)$$

where

$$t_1 = \frac{3-\sqrt{3}}{6} \approx .211 \quad t_2 = \frac{3+\sqrt{3}}{6} \approx .789$$

$$\beta_1 = 1/2 \quad \beta_2 = 1/2$$

The index of each representation is 2, and the roots do strictly interlace.

Example. Consider the same collection of functions discussed in the previous example with the addition of the function $u_4(t) = t^4$. Then

$$C_5 = \{(u_0(t), u_1(t), u_2(t), u_3(t), u_4(t)) \mid t \in [0, 1]\}$$

Let c be the point

$$c = (c_0, c_1, c_2, c_3, c_4) = \left(\int_0^1 1 \, dt, \int_0^1 t \, dt, \int_0^1 t^2 \, dt, \int_0^1 t^3 \, dt, \int_0^1 t^4 \, dt \right) = (1, 1/2, 1/3, 1/4, 1/5).$$

Then $c \in \mathcal{M}_5$ and c is not a boundary point. By Theorem 6.20, the roots of the two principal representations are $\{t_1, t_2, t_3\}$ and $\{s_1, s_2, s_3\}$ where

$$0 = t_1 < t_2 < t_3 < 1 \quad \text{and}$$

$$0 < s_1 < s_2 < s_3 = 1.$$

In each case the index of the representation is $2\frac{1}{2}$. The roots of the representation obtained by Gaussian quadrature are

$$\frac{1}{2} - \frac{\sqrt{15}}{10} \quad \frac{1}{2} \quad \frac{1}{2} + \frac{\sqrt{15}}{10}.$$

Although all representations involve the same number of points, the index of the Gaussian quadrature representation is 3, and therefore it is a canonical representation but not principal.

The following two theorems are stated without proof. The proofs can be found in [28].

Theorem 6.23 *Let c belong to the interior of \mathcal{M}_{n+1} . Consider two different representations of c of index less than or equal to $(n+1)/2$ as follows:*

$$c_i = \sum_{j=1}^p \lambda_j' u_i(t_j') = \sum_{j=1}^q \lambda_j'' u_i(t_j''), \quad i = 0, 1, \dots, n.$$

Then p and q belong to the interval $\left[\left[\frac{n+2}{2}\right], \left[\frac{n+2}{2}\right] + 1\right]$, where $[r]$ denotes the greatest integer $\leq r$. The roots $\{t_j'\}_1^p$ and $\{t_j''\}_1^q$ strictly interlace in (a, b) but may possibly share one or both of the end points a or b .

Theorem 6.24 *Let c be in the interior of \mathcal{M}_{n+1} . For any t^* satisfying $a < t^* < b$, there exists a unique canonical representation of c involving t^* .*

CHAPTER VII

EXPONENTIAL INTERPOLATION

In the method of Gaussian quadrature, the integral of a polynomial p in \mathcal{P}_{n-1} over a closed interval $[a, b]$ is represented by the sum

$$\int_a^b p(t) dt = \sum_{k=1}^n \lambda_k p(t_k)$$

where the nodes, t_1, t_2, \dots, t_n , belong to $[a, b]$, and the weights, $\lambda_1, \lambda_2, \dots, \lambda_n$ are real numbers. Such a representation is exact for all polynomials in \mathcal{P}_{n-1} , and if the nodes are chosen appropriately, it will be exact for all polynomials in \mathcal{P}_{2n-1} . In this chapter, Dirichlet polynomials will be considered. A Dirichlet polynomial is a function f which has the representation

$$f(t) = a_0 e^{\lambda_0 t} + a_1 e^{\lambda_1 t} + \dots + a_n e^{\lambda_n t} \quad (7.1)$$

where a_0, a_1, \dots, a_n are real constants and $\lambda_0, \lambda_1, \dots, \lambda_n$ are complex constants.

Suppose $f : [a, b] \rightarrow \mathbb{R}$ is a function defined by 7.1 and assume $\lambda_0, \lambda_1, \dots, \lambda_n$ are given real constants such that $\lambda_0 < \lambda_1 < \dots < \lambda_n$. Assume also that t_0, t_1, \dots, t_p are given points in $[a, b]$ with $t_0 < t_1 < \dots < t_p$, at which the values of f are known, and let $f(t_k) = \omega_k$ for $k = 0, 1, \dots, p$. Theoretically, if p is large enough, the value of the integral $\int_a^b f(t) dt$ can be determined without knowledge of the coefficients a_0, a_1, \dots, a_n .

Suppose the value of f is known at $n + 1$ distinct points t_0, t_1, \dots, t_n . The set of functions $\{\exp^{\lambda_k t}\}_0^n$ was shown in Chapter VI to be a T-system. So the determinant

$$U = \begin{vmatrix} \exp(\lambda_0 t_0) & \exp(\lambda_0 t_1) & \dots & \exp(\lambda_0 t_n) \\ \exp(\lambda_1 t_0) & \exp(\lambda_1 t_1) & \dots & \exp(\lambda_1 t_n) \\ \vdots & \vdots & & \vdots \\ \exp(\lambda_n t_0) & \exp(\lambda_n t_1) & \dots & \exp(\lambda_n t_n) \end{vmatrix}$$

is positive whenever $a \leq t_0 < t_1 < \dots < t_n \leq b$. Therefore, the coefficients a_0, a_1, \dots, a_n are given by

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = U^{-1} \begin{pmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_n \end{pmatrix}.$$

The integral can be calculated using the computed values of a_0, a_1, \dots, a_n . But if the matrix U is large or poorly conditioned, or if f is not known at $n + 1$ points, then this method of computing the integral is either inefficient or impossible.

Define a vector $c = (c_0, c_1, \dots, c_n)$ in E^{n+1} by

$$c_i = \int_a^b e^{\lambda_i t} dt.$$

Then by Definition 6.8, $c \in \mathcal{M}_{n+1}$. By Theorem 6.16 and Theorem 6.23, there exists a representation of c of the form

$$c_i = \sum_{j=1}^p \gamma_j e^{\lambda_i t_j}, \quad i = 0, 1, \dots, n$$

where $\gamma_1, \gamma_2, \dots, \gamma_p$ are real constants, $a \leq t_1 < t_2 < \dots < t_p \leq b$, and $p \leq \left[\frac{n+2}{2}\right] + 1$. (The notation $[r]$ denotes the greatest integer less than or equal to r .) Then

$$\begin{aligned} \int_a^b f(t) dt &= a_0 \int_a^b e^{\lambda_0 t} dt + a_1 \int_a^b e^{\lambda_1 t} dt + \dots + a_n \int_a^b e^{\lambda_n t} dt \\ &= a_0 \sum_{j=1}^p \gamma_j e^{\lambda_0 t_j} + a_1 \sum_{j=1}^p \gamma_j e^{\lambda_1 t_j} + \dots + a_n \sum_{j=1}^p \gamma_j e^{\lambda_n t_j} \\ &= \gamma_1 \sum_{k=0}^n k e^{\lambda_k t_1} + \gamma_2 \sum_{k=0}^n k e^{\lambda_k t_2} + \dots + \gamma_p \sum_{k=0}^n k e^{\lambda_k t_p} \\ &= \gamma_1 f(t_1) + \gamma_2 f(t_2) + \dots + \gamma_p f(t_p). \end{aligned}$$

For example, let

$$f(t) = a_0 e^{-t} + a_1 e^{-2t} + a_2 e^{-3t} + a_3 e^{-4t}.$$

By making the substitution

$$\begin{aligned} x &= e^{-t} \text{ and} \\ g(x) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3, \end{aligned}$$

the integral of f over the interval $[0, \infty)$ becomes

$$\int_0^\infty f(t) dt = \int_0^1 g(x) dx.$$

As mentioned in the previous chapter, one of the principal representations of the point

$$c = (c_0, c_1, c_2, c_3) = \left(\int_0^1 1 dx, \int_0^1 x dx, \int_0^1 x^2 dx, \int_0^1 x^3 dx \right)$$

is given by

$$c_i = \beta_1 u_i(x_1) + \beta_2 u_i(x_2),$$

where the roots x_1 and x_2 are those found using Gaussian quadrature. When using this method, the orthonormal functions

$$\begin{aligned}\Psi_0 &= 1 \\ \Psi_1 &= \sqrt{3}(2x - 1) \\ \Psi_2 &= \sqrt{5}(6x^2 - 6x + 1).\end{aligned}$$

are computed, and x_1 and x_2 are the following zeros of the polynomial Ψ_2 :

$$\begin{aligned}x_1 &= \frac{3 - \sqrt{3}}{6} \approx .211 \\ x_2 &= \frac{3 + \sqrt{3}}{6} \approx .789.\end{aligned}$$

Using the Lagrange method of calculating the coefficients, we have

$$\begin{aligned}\beta_1 &= \int_0^1 \frac{x - x_2}{x_1 - x_2} dx = \frac{1}{2} \\ \beta_2 &= \int_0^1 \frac{x - x_1}{x_2 - x_1} dx = \frac{1}{2}.\end{aligned}$$

The resulting representation is

$$\begin{aligned}\int_0^1 g(x) dx &= a_0 \int_0^1 1 dx + a_1 \int_0^1 x dx + a_2 \int_0^1 x^2 dx + a_3 \int_0^1 x^3 dx \\ &= \beta_1 g(x_1) + \beta_2 g(x_2).\end{aligned}$$

Since $x = e^{-t}$, then $t = -\ln(x)$ and $g(e^{-t}) = e^t f(t)$. So

$$\int_0^\infty f(t) dt = \frac{1}{2} e^{t_1} f(t_1) + \frac{1}{2} e^{t_2} f(t_2) = \gamma_1 f(t_1) + \gamma_2 f(t_2)$$

where the roots t_1 and t_2 and the weights γ_1 and γ_2 are given by

$$\begin{aligned}t_1 &= -\ln\left(\frac{3 - \sqrt{3}}{6}\right) \\ t_2 &= -\ln\left(\frac{3 + \sqrt{3}}{6}\right) \\ \gamma_1 &= \frac{1}{2} e^{t_1} = \frac{1}{2} \left(\frac{3 - \sqrt{3}}{6}\right)^{-1} \\ \gamma_2 &= \frac{1}{2} e^{t_2} = \frac{1}{2} \left(\frac{3 + \sqrt{3}}{6}\right)^{-1}.\end{aligned}$$

The above procedure is applicable whenever the constants $\lambda_0, \lambda_1, \dots, \lambda_n$ are consecutive negative integers, and it suggests the possibility of applying quadrature to Dirichlet functions which do not have consecutive negative integer constants. For example, let

$$f(t) = a_0 + a_1 e^{-t} + a_2 e^{-4t}.$$

Define functions u_0, u_1 , and u_2 by

$$\begin{aligned} u_0(t) &= 1 \\ u_1(t) &= e^{-t} \\ u_2(t) &= e^{-4t} \end{aligned}$$

A representation of the point

$$c = \left(\int_0^\infty u_0(t) e^{-t} dt, \int_0^\infty u_1(t) e^{-t} dt, \int_0^\infty u_2(t) e^{-t} dt \right)$$

can be found which involves only two roots. As in Gaussian quadrature, the set $\{u_0, u_1, u_2\}$ is used to construct the orthonormal set $\{\Phi_0, \Phi_1, \Phi_2\}$ where

$$\begin{aligned} \Phi_0(t) &= 1 \\ \Phi_1(t) &= \sqrt{3}(2e^{-t} - 1) \\ \Phi_2(t) &= \frac{3}{2}(te^{-4t} - 4e^{-t} + 1). \end{aligned}$$

The zeros of $\Phi_2(t)$ are

$$t_1 = .1951176928 \quad \text{and} \quad t_2 = 1.365272252.$$

Choosing

$$\gamma_1 = .43122554915 \quad \text{and} \quad \gamma_2 = .568774508$$

gives a representation of c of the form

$$c_i = \gamma_1 u_i(t_1) + \gamma_2 u_i(t_2)$$

for $i = 0, 1, 2$. The integral of the function f is given by

$$\int_0^\infty f(t) e^{-t} dt = \gamma_1 f(t_1) + \gamma_2 f(t_2).$$

Other representations can be found by first making the substitutions $x = e^{-t}$ and $g(x) = a_0 + a_1x + a_2x^4$ to get

$$\int_0^\infty f(t)e^{-t} dt = \int_0^1 g(x) dx.$$

Then Gaussian quadrature is applicable. This requires the computation of an orthonormal set $\{\Psi_0, \Psi_1, \Psi_2, \Psi_3\}$, and the zeros of Ψ_3 are the roots required in the representation. The function Ψ_3 is determined to be

$$\Psi_3(x) = \sqrt{7}(20x^3 - 30x^2 + 12x - 1)$$

and the zeros are

$$x_1 = \frac{5 - \sqrt{15}}{10}, \quad x_2 = \frac{1}{2}, \quad x_3 = \frac{5 + \sqrt{15}}{10}.$$

The weights are determined to be

$$\delta_1 = 5/18, \quad \delta_2 = 4/9, \quad \delta_3 = 5/18.$$

Then

$$\begin{aligned} \int_0^\infty f(t)e^{-t} dt &= \int_0^1 g(x) dx \\ &= \delta_1 g(x_1) + \delta_2 g(x_2) + \delta_3 g(x_3) \\ &= \delta_1 f(t_1) + \delta_2 f(t_2) + \delta_3 f(t_3) \end{aligned}$$

where

$$t_1 = -\ln x_1 \approx 2.183011081$$

$$t_2 = -\ln x_2 \approx .6931471806$$

$$t_3 = -\ln x_3 \approx .1195740121.$$

In the above representations, the first required knowledge of the function f only at two points, whereas the second representation, which was found by use of Gaussian quadrature, required three points. The index of the first is 2, and the index of the second is 3. There are two other representations for the integral

$$\int_0^1 g(x) dx$$

which each have index $1\frac{1}{2}$. Each of these is a principal representation. One involves the nodes z_1 and z_2 , where $z_1 = 0$ and $0 < z_2 < 1$. The other involves nodes r_1 and r_2 , where $0 < r_1 < 1$ and $r_2 = 1$. These yield the following representations.

1.

$$\begin{aligned}
\int_0^\infty f(t)e^{-t} dt &= \gamma_1 f(t_1) + \gamma_2 f(t_2) \\
t_1 &= \infty & t_2 &= .3054302439 \\
\gamma_1 &= .3213955956 & \gamma_2 &= .6786044044
\end{aligned}$$

2.

$$\begin{aligned}
\int_0^\infty f(t)e^{-t} dt &= \delta_1 f(s_1) + \delta_2 f(s_2) \\
s_1 &= .9432378690 & s_2 &= 0 \\
\delta_1 &= .8188198595 & \delta_2 &= .1811801405
\end{aligned}$$

In general, it would be desirable to find an efficient way of representing Dirichlet polynomials which have exponents with arbitrary negative coefficients. For example consider the function f given by

$$f(t) = e^{-2t} + e^{-t} + e^{-\frac{1}{2}t} + e^{-\frac{1}{3}t}.$$

The functions e^{-2t} , e^{-t} , $e^{-\frac{1}{2}t}$ and $e^{-\frac{1}{3}t}$ form a T-System. So there should exist constants $\gamma_1, \gamma_2, \gamma_3$ and nodes t_1, t_2, t_3 such that

$$\int_0^\infty f(t) \cdot e^{-t} dt = \gamma_1 f(t_1) + \gamma_2 f(t_2) + \gamma_3 f(t_3).$$

CHAPTER VIII

QUADRATURE METHODS

Let A be an $n \times n$ constant matrix, c a constant vector in \mathbb{R}^n , and I an interval. Let x be a function from I to \mathbb{R}^n . Consider the linear system

$$\dot{x} = Ax \quad y = c^T x \quad x(t_0) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \quad t_0 \in I \quad (8.1)$$

It can be assumed that $t_0 = 0$, since an appropriate translation of the variable transforms any system into this form. Assume that I has the form $[0, \infty)$ or $[0, b]$ for a real number $b > 0$. Assume also that the system is discretely observable with respect to the set of points $\{t_1, t_2, \dots, t_m\} \subset I$.

The solution for x in (1) is given by

$$x(t) = \exp(At)x(0)$$

The $n \times n$ matrix $\exp(At)$ has the form $\exp(At) = (q_{ij}(t))$, where each $q_{ij}(t)$ is given by

$$q_{ij}(t) = \sum_{k=1}^s p_{ijk}(t)e^{\lambda_k t}$$

where $\{\lambda_1, \lambda_2, \dots, \lambda_s\}$ is the set of distinct eigenvalues of A , and each p_{ijk} is a polynomial of degree less than n .

Then

$$y(t) = c^T x(t) = c^T \exp(At)x(0) = \sum_{i,j=1}^n b_{ij} q_{ij}(t)$$

where each b_{ij} is a constant. Regrouping the terms gives

$$y(t) = \sum_{i=1}^s r_i(t)e^{\lambda_i t},$$

where each r_i is a polynomial of degree less than n .

The solution y can be determined directly from knowledge of the vectors c and $x(0)$. But it can also be determined from known values of y at points t_1, t_2, \dots, t_m .

Suppose there exists an invertible constant matrix Q such that

$$y(t) = c^T \exp(At) Q Q^{-1} x(0)$$

and

$$c^T \exp(At) Q = \begin{pmatrix} \phi_1(t) & \phi_2(t) & \cdots & \phi_n(t) \end{pmatrix}$$

where $\{\phi_1(t), \phi_2(t), \dots, \phi_n(t)\}$ is an orthonormal collection of functions on the interval I with respect to the inner product

$$\langle f, g \rangle = \int_I f(t)g(t)d\mu(t). \quad (8.2)$$

Then, letting

$$Q^{-1}x(0) = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix},$$

y has the representation

$$y = \omega_1 \phi_1(t) + \omega_2 \phi_2(t) + \cdots + \omega_n \phi_n(t)$$

on I . Then for each k , the constant ω_k is determined without knowledge of Q or $x(0)$ if $\langle y, \phi_k \rangle$ is known for each k . These values can be determined from the known values $y(t_1), y(t_2), \dots, y(t_m)$ if a quadrature formula involving points t_1, t_2, \dots, t_m can be found which is exact for all $\langle q_{ij}, \phi_k \rangle$, where $i, j, k = 1, 2, \dots, n$.

If such a quadrature formula can be found, then for each $k = 1, 2, \dots, m$,

$$\begin{aligned} \omega_k &= \langle y, \phi_k \rangle \\ &= \int y(t) \phi_k(t) d\mu(t) \\ &= A_1 y(t_1) \phi_k(t_1) + A_2 y(t_2) \phi_k(t_2) + \cdots + A_n y(t_m) \phi_k(t_m), \end{aligned} \quad (8.3)$$

The eigenvalues of the matrix A will determine the form of the solution y , and it will be assumed that A is in Jordan canonical form. This assumption leads to no loss of generality since every matrix is similar to a matrix J in Jordan canonical form, and the linearity of the dynamical system is preserved under a change of variables of the form $Z = Px$, where $A = P^{-1}JP$.

8.1 Nilpotent Matrices

If A is a nilpotent matrix, then the solution is a polynomial and can be written as a linear combination of orthogonal polynomials. In special cases, these orthogonal polynomials have been computed. In general, they can be obtained using a three term recursion formula. Gaussian quadrature is applicable, and formulas are available for some particular intervals of integration and weight functions. See [8].

8.2 Matrices With One Real Eigenvalue

Let A be the following $n \times n$ matrix in Jordan form:

$$A = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

Then the solution y of the linear system (1) is given by

$$y(t) = c^T x(t) = c^T \exp(At)x(0),$$

where $\exp(At)$ is the matrix

$$\exp(At) = \exp(\lambda t)M(t),$$

with

$$M(t) = \begin{pmatrix} 1 & t & t^2/2 & t^3/3! & \dots & t^{n-1}/(n-1)! \\ & 1 & t & t^2/2 & \dots & t^{n-2}/(n-2)! \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & t & t^2/2 \\ & & & & 1 & t \\ & & & & & 1 \end{pmatrix}$$

Suppose there exists a set $\{\phi_k\}_0^{n-1}$ of polynomials which are orthonormal with respect to the inner product

$$\langle f, g \rangle = \int_I f(t)g(t)dt,$$

where

$$\begin{aligned}
 \phi_0(t) &= s_{11} \\
 \phi_1(t) &= s_{21} + s_{22}t \\
 \phi_2(t) &= s_{31} + s_{32}t + s_{33}t^2 \\
 &\vdots \\
 \phi_{n-1}(t) &= s_{n1} + s_{n2}t + \cdots + s_{nn}t^{n-1},
 \end{aligned}$$

and for each $k = 1, 2, \dots, n$, $s_{kk} \neq 0$. See [19] for existence of such a set and recursion formulas for computation of the polynomials.

Let $c^T = (c_1 \ c_2 \ \cdots \ c_n)$. Since the system is observable, $c_k \neq 0$ for each $k = 1, 2, \dots, n$. Therefore the matrix

$$C = \begin{pmatrix} c_1 & c_2 & c_3 & c_4 & \cdots & c_n \\ & c_1 & c_2 & c_3 & \cdots & c_{n-1} \\ & & \frac{c_1}{2} & \frac{c_2}{2} & \cdots & \frac{c_{n-2}}{2} \\ & & & \frac{c_1}{3!} & \cdots & \frac{c_{n-3}}{3!} \\ & & & & \ddots & \vdots \\ & & & & & \frac{c_1}{(n-1)!} \end{pmatrix}$$

is invertible. Let

$$S = \begin{pmatrix} s_{11} & s_{21} & s_{31} & s_{41} & \cdots & s_{n1} \\ & s_{22} & s_{32} & s_{42} & \cdots & s_{n2} \\ & & s_{33} & s_{43} & \cdots & s_{n3} \\ & & & s_{44} & \cdots & s_{n4} \\ & & & & \ddots & \vdots \\ & & & & & s_{nn} \end{pmatrix}.$$

Then there exists a matrix

$$Q = \begin{pmatrix} q_{11} & q_{12} & q_{13} & q_{14} & \cdots & q_{1n} \\ & q_{22} & q_{23} & q_{24} & \cdots & q_{2n} \\ & & q_{33} & q_{34} & \cdots & q_{3n} \\ & & & q_{44} & \cdots & q_{4n} \\ & & & & \ddots & \vdots \\ & & & & & q_{nn} \end{pmatrix}$$

satisfying $CQ = S$. Since each $s_{kk} \neq 0$, Q is invertible, and

$$\begin{aligned} c^T \exp(At)Q &= c^T \exp(\lambda t)M(t)Q \\ &= \exp(\lambda t) \begin{pmatrix} 1 & t & t^2 & \cdots & t^{n-1} \end{pmatrix} CQ \\ &= \exp(\lambda t) \begin{pmatrix} 1 & t & t^2 & \cdots & t^{n-1} \end{pmatrix} S \\ &= \exp(\lambda t) \begin{pmatrix} \phi_0(t) & \phi_1(t) & \cdots & \phi_{n-1}(t) \end{pmatrix} \end{aligned}$$

Let

$$Q^{-1}x(0) = \begin{pmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_{n-1} \end{pmatrix}.$$

Then the representation for y is

$$\begin{aligned} y(t) &= c^T \exp(At)x(0) = c^T \exp(At)QQ^{-1}x(0) \\ &= \exp(\lambda t)(\omega_0\phi_0(t) + \omega_1\phi_1(t) + \cdots + \omega_{n-1}\phi_{n-1}(t)). \end{aligned}$$

Let f be defined by $f(t) = \exp(-\lambda t)y(t)$. Since the set $\{\phi_k\}_0^{n-1}$ is orthonormal, each coefficient ω_k is given by

$$\omega_k = \langle f, \phi_k \rangle.$$

Since f and ϕ_k are polynomials of degree less than or equal to $n-1$, then $f\phi_k$ is the product of $\exp(\lambda t)$ and a polynomial of degree less than or equal to $2n-2$. The functions $\exp(\lambda t)$, $t\exp(\lambda t)$, $t^2\exp(\lambda t)$, \dots , $t^{2n-2}\exp(\lambda t)$ form a T-System. So there exists a quadrature formula involving n points t_1, t_2, \dots, t_n in I and weights a_1, a_2, \dots, a_n . If the values $y(t_1), y(t_2), \dots, y(t_n)$ are known, ω_k is given by

$$\begin{aligned} \omega_k &= a_1 \exp(-\lambda t_1)y(t_1)\phi_k(t_1) + \\ &\quad a_2 \exp(-\lambda t_2)y(t_2)\phi_k(t_2) + \cdots + \\ &\quad a_n \exp(-\lambda t_n)y(t_n)\phi_k(t_n). \end{aligned}$$

8.3 Distinct Real Eigenvalues

Let A be an $n \times n$ matrix with distinct real eigenvalues, and consider the linear system (8.1). Assume $t_0 = 0$, c has no zero components, and

$$A = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix},$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_n$. The solution for y is given by

$$y(t) = c^T \text{diag}\{\exp(\lambda_1 t), \exp(\lambda_2 t), \dots, \exp(\lambda_n t)\} x(0),$$

which has the form

$$y(t) = b_1 \exp(\lambda_1 t) + b_2 \exp(\lambda_2 t) + \cdots + b_n \exp(\lambda_n t)$$

for real constants b_1, b_2, \dots, b_n .

Using the Gram-Schmidt orthonormalizing process, an orthonormal set of functions $\{\phi_1, \phi_2, \dots, \phi_n\}$ with respect to the inner product (8.2) can be found satisfying

$$\begin{aligned} \phi_1(t) &= s_{11} e^{\lambda_1 t} \\ \phi_2(t) &= s_{21} e^{\lambda_1 t} + s_{22} e^{\lambda_2 t} \\ &\vdots \\ \phi_n(t) &= s_{n1} e^{\lambda_1 t} + s_{n2} e^{\lambda_2 t} + \cdots + s_{nn} e^{\lambda_n t} \end{aligned}$$

where $s_{kk} \neq 0$ for each $k = 1, \dots, n$. Let

$$c^T = (c_1, c_2, \dots, c_n),$$

$$S = \begin{pmatrix} s_{11} & s_{21} & s_{31} & \cdots & s_{n1} \\ & s_{22} & s_{32} & \cdots & s_{n2} \\ & & s_{33} & \cdots & s_{n3} \\ & & & \cdots & \\ & & & & s_{nn} \end{pmatrix}$$

and $Q = \text{diag}\{1/c_1, 1/c_2, \dots, 1/c_n\} S$.

Then

$$\begin{aligned}
 y(t) &= c^T \text{diag}\{\exp(\lambda_1 t), \exp(\lambda_2 t), \dots, \exp(\lambda_n t)\} x(0) \\
 &= \begin{pmatrix} \exp(\lambda_1 t) & \exp(\lambda_2 t) & \cdots & \exp(\lambda_n t) \end{pmatrix} \text{diag}\{c_1, c_2, \dots, c_n\} x(0) \\
 &= \begin{pmatrix} \exp(\lambda_1 t) & \exp(\lambda_2 t) & \cdots & \exp(\lambda_n t) \end{pmatrix} \text{diag}\{c_1, c_2, \dots, c_n\} Q Q^{-1} x(0).
 \end{aligned}$$

Note that

$$\begin{aligned}
 \begin{pmatrix} \exp(\lambda_1 t) & \exp(\lambda_2 t) & \cdots & \exp(\lambda_n t) \end{pmatrix} \text{diag}\{c_1, c_2, \dots, c_n\} Q &= \\
 \begin{pmatrix} \exp(\lambda_1 t) & \exp(\lambda_2 t) & \cdots & \exp(\lambda_n t) \end{pmatrix} S &= \\
 \begin{pmatrix} \phi_1(t) & \phi_2(t) & \cdots & \phi_n(t) \end{pmatrix}. &
 \end{aligned}$$

Denote the constant matrix $Q^{-1} x(0)$ by $\begin{pmatrix} \omega_1 & \omega_2 & \cdots & \omega_n \end{pmatrix}$. Then

$$\begin{aligned}
 y(t) &= \begin{pmatrix} \phi_1(t) & \phi_2(t) & \cdots & \phi_n(t) \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix} \\
 &= \omega_1 \phi_1(t) + \omega_2 \phi_2(t) + \cdots + \omega_n \phi_n(t). \tag{8.4}
 \end{aligned}$$

Since $\{\phi_1, \phi_2, \dots, \phi_n\}$ is an orthonormal set, the coefficients $\omega_1, \omega_2, \dots, \omega_n$ in (8.4) are given by

$$\omega_k = \int_I y(t) \phi_k(t) d\mu(t) \quad k = 1, 2, \dots, n.$$

Since $y(t)$ and $\phi_k(t)$ are linear combinations of the functions $e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_n t}$, then $y(t)\phi_k(t)$ is a linear combination of functions of the form $e^{(\lambda_i + \lambda_j)t}$. Let $B = \{\beta_1, \beta_2, \dots, \beta_s\}$ be the set of distinct elements of $\{\lambda_i + \lambda_j \mid i, j = 1, 2, \dots, n\}$, where $2n - 1 \leq s \leq \frac{n(n+1)}{2}$. Then for each $k = 1, 2, \dots, n$, there exist constants $A_{k1}, A_{k2}, \dots, A_{ks}$ such that

$$y(t)\phi_k(t) = A_{k1}e^{\beta_1 t} + A_{k2}e^{\beta_2 t} + \cdots + A_{ks}e^{\beta_s t}.$$

Then

$$\omega_k = \int_I (A_{k1}e^{\beta_1 t} + A_{k2}e^{\beta_2 t} + \cdots + A_{ks}e^{\beta_s t}) d\mu(t).$$

Suppose I is the closed finite interval $[a, b]$. The functions $\exp(\beta_1 t), \exp(\beta_2 t), \dots, \exp(\beta_s t)$ form a Tchebycheff system (T-system) over $[a, b]$. The point $c = (c_1, c_2, \dots, c_s)$,

where

$$c_i = \int_a^b \exp(\beta_i t) d\mu(t),$$

belongs to the moment space \mathcal{M}_s . So by Theorems 6.16 and 6.19, there exists a quadrature formula involving a set of r points t_1, t_2, \dots, t_r , where $r \leq \frac{s+2}{2}$. If any of the points are preassigned, a quadrature formula still exists, but it might require more than $\frac{s+2}{2}$ points. The maximum number of points required would be s .

It is shown in [7] that if a quadrature rule of the form

$$\int_a^b \exp(\beta_i t) d\mu(t) = \sum_{j=1}^n \alpha_j \exp(\beta_i t_j) \quad i = 1, 2, \dots, s$$

exists, then $s \leq 2n$. Since $2n - 1 \leq s$, we have $s = 2n - 1$ or $s = 2n$. If, however, $s > 2n$, the values of $y(t)$ at more than n points would be required.

8.4 Real Repeated Eigenvalues

Let A be an $n \times n$ matrix with real eigenvalues, where some of the eigenvalues are repeated. Let $\lambda_1, \lambda_2, \dots, \lambda_s$ be the distinct eigenvalues. Then the solution y of (8.1) has the form

$$y(t) = \sum_{i=1}^s r_i(t) \exp(\lambda_i t)$$

where each r_i is a polynomial of degree less than n . This can be written as a linear combination of the linearly independent functions:

$$\begin{aligned} & \exp(\lambda_1 t), \quad t \exp(\lambda_1 t), \quad t^2 \exp(\lambda_1 t), \quad \dots, \quad t^{m_1} \exp(\lambda_1 t), \\ & \exp(\lambda_2 t), \quad t \exp(\lambda_2 t), \quad t^2 \exp(\lambda_2 t), \quad \dots, \quad t^{m_2} \exp(\lambda_2 t), \\ & \dots \\ & \exp(\lambda_s t), \quad t \exp(\lambda_s t), \quad t^2 \exp(\lambda_s t), \quad \dots, \quad t^{m_s} \exp(\lambda_s t) \end{aligned} \quad (8.5)$$

where $m_i + 1$ is the multiplicity of the eigenvalue λ_i .

Theoretically, this collection of functions can be orthonormalized, and the previous method could be used to determine a solution of (8.1) by evaluating constants of the form

$$\omega_k = \int_I y(t) \phi_k(t) d\mu(t).$$

The function $y(t)\phi_k(t)$ would be in the span of a set T , where T is a linearly independent subset of a set

$$S = \bigcup_{k=1}^m \{ \exp(\alpha_k t) \cos \beta_k t, \exp(\alpha_k t) \sin \beta_k t, \\ t \exp(\alpha_k t) \cos \beta_k t, t \exp(\alpha_k t) \sin \beta_k t, \\ \dots, \\ t^{m_k} \exp(\alpha_k t) \cos \beta_k t, t^{m_k} \exp(\alpha_k t) \sin \beta_k t \}.$$

Quadrature formulas for such integrals are known for special cases. See, for example, [15] and [30]. The existence of a quadrature formula is guaranteed in case S is a T-system with respect to the interval of integration. For example, the system of $2m + 1$ functions

$$1, \cos t, \sin t, \dots, \cos mt, \sin mt$$

is a T-system on any interval $[a, b]$ of length less than 2π . Furthermore, if $\{u_i\}$ is a T-system and the function r is positive and continuous, then $\{ru_i\}$ is a T-system. See [28].

Other examples of T-systems are eigenfunctions of Sturm-Liouville operators. Let the operator L be defined by

$$L(\phi) = -\frac{d}{dx} \left(p \frac{d\phi}{dx} \right) + q\phi$$

where p is continuous and positive on $[a, b]$. Let $K(x, y)$ be the Green's function associated with the eigenvalue problem

$$L(\phi) = \lambda\phi$$

with boundary conditions

$$\begin{aligned} \phi(a) \sin \alpha - p(a) \phi'(a) \cos \alpha &= 0 \\ \phi(b) \sin \beta + p(b) \phi'(b) \cos \beta &= 0. \end{aligned}$$

If $K(x, y)$ satisfies certain conditions, then the set of eigenfunctions $\phi_0, \phi_1, \dots, \phi_n$ is a T-system on any closed subinterval of (a, b) . See [10].

8.6 Remarks

In summary, the determination of a solution y to (1) can be found by a method which depends on orthonormalizing a given set of functions which form a T-system and finding the nodes and weights in a quadrature formula. In some cases, this method has been completely worked out. In the remaining cases, the theory of T-systems guarantees the existence of quadrature formulas, but the number of data points required might be greater than n . Further study of these remaining cases to determine the least number of data points required and efficient methods of determining weights and nodes would be helpful.

REFERENCES

- [1] N. I. Akhiezer. *The Classical Moment Problem*. Hafner Publishing Company, New York, 1965.
- [2] Linda Allen, Truman Lewis, Clyde F. Martin, and Mark Stamp. A mathematical analysis and simulation of a localized measles epidemic. *Applied Mathematics and Computation*, 39:61–77, 1990.
- [3] G. S. Ammar, W. B. Gragg, and L. Reichel. Determination of Pisarenko frequency estimates as eigenvalues of an orthogonal matrix. In *Advanced Algorithms and Architectures for Signal Processing II*, SPIE Vol. 826, August 1987.
- [4] David L. Barrow. On multiple node Gaussian quadrature formulae. *Mathematics of Computation*, 32(142), April 1978.
- [5] Franca Calìò, Walter Gautschi, and Elena Marchetti. On computing Gauss-Kronrod quadrature formulae. *Mathematics of Computation*, 47(176), October 1986.
- [6] Man-Duen Choi. Tricks or treats with the Hilbert matrix. *American Mathematical Monthly*, May 1983.
- [7] M. Cordero-Vourtsanis, C. Martin, and J. Miller. *Gaussian Quadrature for Products of Exponential Functions*. Technical Report 149, Texas Tech University, 1990. Center for Applied Systems Analysis.
- [8] Philip J. Davis. *Interpolation and Approximation*. Dover Publications, Inc., New York, 1975.
- [9] Philip J. Davis and Philip Rabinowitz. *Numerical Integration*. Blaisdell Publishing Company, Waltham, Massachusetts, 1967.
- [10] F. R. Gantmacher and M. G. Krein. *Oscillation Matrices and Kernels and Small Oscillations of Mechanical Systems*. GITTL, Moscow, 2 edition, 1950.
- [11] Walter Gautschi. Construction of Gauss-Christoffel quadrature formulas. *Mathematics of Computation*, 22:251–270, 1968.
- [12] Walter Gautschi. On the construction of Gaussian quadrature rules from modified moments. *Mathematics of Computation*, 24(110), April 1970.
- [13] Walter Gautschi. A survey of Gauss-Christoffel quadrature formulae. In *E. B. Christoffel*, pages 72–147, Birkhauser, 1981.

- [14] D. Gilliam, J. Lund, and C. Martin. Gaussian quadrature for trigonometric polynomials: an example from observability. *Applicable Analysis*, 44:183–189, 1992.
- [15] D. S. Gilliam, J. R. Lund, and C. F. Martin. Inverse parabolic problems and discrete orthogonality. *Numerische Mathematik*, 59:361–383, 1991.
- [16] D. S. Gilliam, B. A. Mair, and C. F. Martin. Determination of initial states of parabolic systems from discrete data. *Inverse Problems*, 6:737–747, 1990.
- [17] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, 2 edition, 1989.
- [18] Gene H. Golub and John H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23:221–230, 1969.
- [19] William B. Gragg. Matrix interpretations and applications of the continued fraction algorithm. *Rocky Mountain Journal of Mathematics*, 4(2), Spring 1974.
- [20] William B. Gragg and William J. Harrod. The numerically stable reconstruction of Jacobi matrices from spectral data. *Numerische Mathematik*, 44, 1984.
- [21] Preston C. Hammer and Howard H. Wicke. Quadrature formulas involving derivatives of the integrand. *Mathematics of Computation*, 14:3–7, 1960.
- [22] C. G. Harris and W. A. B. Evans. Extension of numerical quadrature formulae to cater for end point singular behaviours over finite intervals. *Inter. J. Computer Math.*, 6, 1977, Section B.
- [23] Robert Hermann. A lie-theoretic setting for the classical interpolation theories. *Acta Applicandae Mathematicae*, 4, 1986.
- [24] T. H. Hildebrandt. *Introduction to the Theory of Integration*. Academic Press, New York, 1963.
- [25] Alberto Isidori. *Nonlinear Control Systems, 2d Ed.* Springer-Verlag, Berlin, 1989.
- [26] Samuel Karlin and Allan Pinkus. An extremal property of multiple Gaussian nodes. In *Studies in Spline Functions and Approximation Theory*, Academic Press, New York, 1976.
- [27] Samuel Karlin and Allan Pinkus. Gaussian quadrature formulae with multiple nodes. In *Studies in Spline Functions and Approximation Theory*, Academic Press, New York, 1976.

- [28] Samuel Karlin and William J. Studden. *Tchebycheff Systems: with Applications in Analysis and Statistics*. Interscience Publishers, New York, 1966.
- [29] Sven-Åke Gustafson. On computational applications of the theory of moment problems. *Rocky Mountain Journal of Mathematics*, 4(2), Spring 1974.
- [30] C. J. Knight and A. C. R. Newbery. Trigonometric and Gaussian quadrature. *Mathematics of Computation*, 24(111), July 1970.
- [31] Peter Lancaster and Miron Tismenetsky. *The Theory of Matrices*. Academic Press, Orlando, Florida, 1985.
- [32] Clyde Martin and Ilias Iakovidis. The numerical stability of observability. In *Realization and Modeling in System Theory*, Proceedings of the International Symposium, NTMS. M. KAASHOEK, 1990.
- [33] Clyde Martin, Mark Stamp, and Ziaochang Wang. Discrete observability and numerical quadrature. *IEEE Trans. Aut. Control*, 36:1337–1340, November 1991.
- [34] Clyde Martin and Shishen Xie. Observability of electropotential on nested cylindrical domain, part i: the continuous solution. *Applied Mathematics and Computation*, 39:207–244.
- [35] Clyde Martin and Shishen Xie. Observability of electropotential on nested cylindrical domain, part ii: numerical solution. *Applied Mathematics and Computation*, 39:245–270.
- [36] C. A. Micchelli and T. J. Rivlin. Quadrature formulae and hermite-birkhoff interpolation. *Advances in Mathematics*, 11, 1973.
- [37] James M. Ortega. *Matrix Theory: A Second Course*. Plenum Press, New York, 1987.
- [38] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1964.
- [39] Jennifer K. Smith. *The Mathematics of Interpolation and Sampling*. Master's thesis, Texas Tech University, 1986.
- [40] Jan Van Tiel. *Convex Analysis*. John Wiley & Sons, New York, 1984.
- [41] P. Turán. On the theory of the mechanical quadrature. *Acta. Sci. Math., Szeged. 12, Par. A*, 30–37, 1950.
- [42] H. S. Wall. *Analytic Theory of Continued Fractions*. D. Van Nostrand Company, Inc., New York, 1948.

- [43] John C. Wheeler. Modified moments and gaussian quadratures. *Rocky Mountain Journal of Mathematics*, 4(2), Spring 1974.
- [44] W. J. Wiscombe and J. W. Evans. Exponential-sum fitting of radiative transmission functions. *Journal of Computational Physics*, 24, 1977.